0UNIVERSIDADE ESTADUAL DE MARINGÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO

RODRIGO IVÁN CONTRERAS SOTO

Estrutura genética e associação genômica baseada em haplótipos para caracteres agronômicos em soja

MARINGÁ
PARANÁ – BRASIL
MAIO - 2017

RODRIGO IVÁN CONTRERAS SOTO

Estrutura genética e associação genômica baseada em haplótipos para caracteres agronômicos em soja

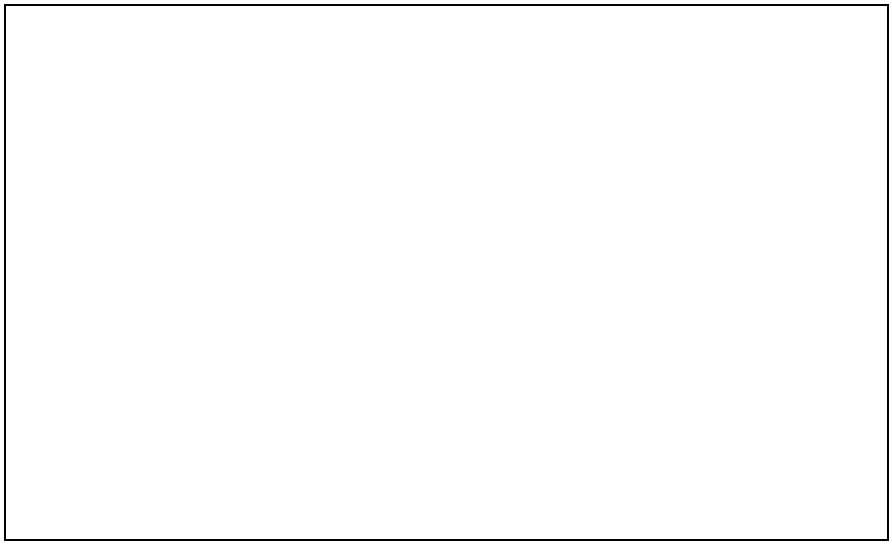
Tese apresentada à Universidade Estadual de Maringá, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para a obtenção do Título de Doutor.

Orientador: Prof. Dr. Carlos Alberto Scapim.

MARINGÁ
PARANÁ – BRASIL
MAIO - 2017

FICHA CATALOGRÁFICA.
É ELABORADA PELA BIBLIOTECA CENTRAL DA UEM E DEVERÁ SER IMPRESSA NO VERSO DA FOLHA DE ROSTO, OU SEJA, NO VERSO DA FOLHA ANTERIOR.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM. Maringá, PR, Brasil)



PÁGINA DESTINADA À FOLHA DE APROVAÇÃO. ENCONTRA-SE NA SECRETARIA DO PGM

Aos meus pais, Maximo Contreras e Margarita Soto.

Às minhas irmãs, Maggy e Kathy.

À minha namorada, Carolina Bertuzzi.

A todos os professores e amigos que sempre me apoiaram.

Ao professor doutor Freddy Mora e ao Dr. Ariel Salvatierra.

Com carinho, dedico.

AGRADECIMENTOS

A Deus, por todas as graças, principalmente por ter me concedido a vida, minha família, meus amigos e muitas bênçãos.

À Universidade Estadual de Maringá (UEM) e ao Programa de Pós-Graduação em Genética e Melhoramento (PGM), pela oportunidade para a realização do Curso de Doutorado.

À Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), pela concessão da bolsa de estudos e pelo apoio financeiro.

À Cooperativa Central de Pesquisa Agrícola (Coodetec), pela disponibilidade dos dados e aos seus pesquisadores e funcionários, que contribuíram imensamente para o desenvolvimento do projeto.

Aos professores doutores Carlos Alberto Scapim e Ivan Schuster, pelos ensinamentos, disponibilidade, dedicação, sugestões e confiança no meu trabalho.

Aos Coorientadores, professores doutores Ronald Barth Pinto e Hugo Zeni, pelos ensinamentos, comentários e sugestões.

Aos Secretários do Programa de Pós-Graduação em Genética e Melhoramento, Maria Valquíria Magro e Francisco José da Cruz, pelo apoio e favores prestados.

Aos colegas do grupo de milhos especiais, em especial aos amigos Marlon Coan, Diego Ary Rizzardi, Alex Sandro Torre Figueiredo, Alessandra Guedes Baleroni e Camila Castro, por toda a colaboração e companheirismo.

BIOGRAFIA

RODRIGO IVÁN CONTRERAS SOTO, filho de Maximo Contreras e Margarita Soto, nasceu no dia 25 de Janeiro de 1988, na cidade de La Laja, Oitava Região, Chile.

Em dezembro de 2001, concluiu o Ensino Fundamental, na Escola Estadual Miguel Luis Amunátegui, na cidade de San Rosendo, Oitava Região, Chile. Concluiu o Ensino Médio, em dezembro de 2005, no Liceo Estadual Politécnico Héroes de la Concepción, na cidade de La Laja, Oitava Região, Chile.

Ingressou no Curso de Engenharia em Biotecnologia Vegetal, em março de 2006, na Universidad de Concepción (UDEC), cidade de Concepción, Oitava Região, Chile, obtendo o título de Engenheiro em Biotecnologia em março de 2011.

Em março de 2011, ingressou no Curso de Mestrado do Programa de Pós-Graduação em Ciências Florestais, da Universidad de Concepción (UDEC), na cidade de Concepción, Oitava Região, Chile, obtendo o título de Mestre em outubro de 2013.

Em fevereiro de 2014, ingressou no Curso de Doutorado do Programa de Pós-Graduação em Genética e Melhoramento (PGM), da Universidade Estadual de Maringá (UEM), Maringá, Paraná, Brasil.

SUMÁRIO

RESUMO

CONTRERAS-SOTO, Rodrigo Iván, D.Sc. Universidade Estadual de Maringá, maio, 2017. **Estrutura genética e associação genômica baseada em haplótipos para caracteres agronômicos em soja**. Orientador: Carlos Alberto Scapim. Coorientadores: Ivan Schuster e Ronald José Barth Pinto.

A soja (*Glycine max* L.) é uma espécie autógama, cuja base genética no Brasil é resultado de vários ciclos de seleção e recombinações entre um reduzido número de genótipos selecionados a partir de cultivares dos Estados Unidos. Este frequente processo de seleção, mistura de populações e cruzamentos de um número reduzido de cultivares poderia conduzir a um incremento no relacionamento genético e afetar os padrões da estrutura populacional. Estes fatores influenciam os padrões dos blocos de desequilíbrio de ligação e podem servir como uma abordagem, visando à busca de *loci* associados a caracteres de importância agronômica em cultivares de soja tropical. O mapeamento de *loci* de caracteres quantitativos por meio do uso do desequilíbrio de ligação provê uma valiosa abordagem para o estudo da base genética de caracteres complexos em soja. O mapeamento por associação de genoma amplo baseado em haplótipos tem sido proposto como uma abordagem complementar que intensifica os benefícios do desequilíbrio de ligação e que permite avaliar os determinantes genéticos de caracteres agronômicos complexos. Os objetivos do presente trabalho foram: analisar os blocos em desequilíbrio de ligação, estimar a estrutura populacional e o relacionamento genético, por meio da genotipagem com a plataforma iScan BARCSoy6K da Illumina, em 141 cultivares de soja tropical. O mapeamento por associação de genoma amplo (GWAS) foi utilizado para identificar regiões genômicas que controlam o peso de 100 sementes (SW), altura da planta (PH), rendimento de grãos (SY) e caracteres de floração (dias para floração e maturação, DTF e DTM, respectivamente) em um painel de mapeamento de associação de soja, usando marcadores de polimorfismo de um único nucleotídeo (SNP) e informação de haplótipos. As cultivares de soja (N=141) foram avaliadas em cinco locais do Sul do Brasil, totalizando oito ambientes. Os resultados revelaram fortes correlações positivas e negativas entre floração e maturidade com caracteres de rendimento de grãos e significativas associações, que representam trinta e três, vinte e nove, cinquenta e sete, setenta e dois e quarenta haplótipos baseados em SNPs associados com SY, SW, PH, DTM e DTF, respectivamente, em

dois ou mais ambientes. Especificamente, GWAS baseada em haplótipos identificou três haplótipos significativamente coassociados entre DTF, DTM e caracteres relacionados ao rendimento em diferentes e específicos ambientes. Estes resultados sugerem que estas regiões genômicas poderiam conter genes com efeitos pleiotrópicos, controlando caracteres relacionados ao rendimento e ao tempo para floração e maturidade, e estão intimamente ligados a outros múltiplos genes com altas taxas de desequilíbrio de ligação.

Palavras-chave: Haplótipos, efeitos pleiotrópicos, BARKSoy6K, soja tropical.

ABSTRACT

CONTRERAS-SOTO, Rodrigo Iván, D.Sc. Universidade Estadual de Maringá, may, 2017. **Population structure and genomic association based on haplotypes for agronomic traits in soybean.** Adviser: Carlos Alberto Scapim. Committee Members: Ivan Schuster and Ronald José Barth Pinto.

Soybean (*Glycine max* L.) is an annual, self-pollinated species, whose genetic base in Brazil is the result of several cycles of selection and effective recombination among a relatively small number of selected genotypes from the USA cultivars. This frequent selection, admixed population and the crossing of a small number of cultivars can lead to increase the genetic relationship and affect the patterns of population complementary approach to intensify benefits from LD, which enable to assess the genetic determinants of agronomic traits. Thus the objectives of this research were to analyze LD blocks, estimate population structure and relatedness through of genotyping of 141 cultivars of tropical soybean by using a BARCSoy6K of Illumina iScan platform. The GWAS was undertaken to identify genomic regions that control 100-seed weight (SW), plant height (PH), seed yield (SY) and flowering traits (Days to flowering and maturity, DTF and DTM, respectively) in a soybean association mapping panel using single nucleotide polymorphism (SNP) markers and haplotype information. The soybean cultivars (N=141) were field-evaluated across five locations of southern Brazil, eight environments. Our results revealed strong positive and negative correlations of flowering and maturity with yield-related traits and the significant association of thirty-three, twenty-nine, fifty-seven, seventy-two and forty SNP-based haplotypes with SY, SW, PH, DTM and DTF, respectively, in two or more environments. Specifically, haplotype-based GWAS identified three haplotypes (Gm12_Hap12; Gm19_Hap42 and Gm20_Hap32) significantly co-associated with DTF, DTM and yield-related traits in specific and multiple environments. These results indicate that these genomic regions may contain genes with pleiotropic effects controlling traits related to yield and time to flowering and maturity and are tightly linked with others multiple genes with high rates of linkage disequilibrium.

Keywords: Haplotypes; pleiotropic effects; BARCSoy6K; tropical soybean.

1. INTRODUÇÃO GERAL

A soja é uma das culturas mais comercializadas no mundo e a que mais cresceu em área semeada e em produtividade, no Brasil, nas últimas três décadas. A soja corresponde atualmente a 52% da área semeada em grãos do país, com uma previsão da taxa de crescimento anual de produção de 2,43% até 2019, próxima da taxa mundial, estimada em 2,56% para os próximos 10 anos (CONAB, 2017).

O aumento da produtividade da cultura da soja no Brasil está associado aos avanços tecnológicos, ao manejo e à eficiência dos produtores, principalmente aos resultados obtidos pelas pesquisas de organismos públicos e empresas privadas, que em conjunto geraram programas de melhoramento genético que possibilitaram o incremento da produtividade média, atingindo os maiores índices mundiais (Sediyama, 2015).

Os avanços da biotecnologia, por meio do uso dos marcadores moleculares, permitiram avaliar a diversidade genética da cultura no Brasil. Embora exista um grande número de cultivares de soja no país, há pouca variabilidade genética entre elas, principalmente por serem originárias de poucas linhagens ancestrais, o que resulta em uma base genética estreita (Priolli et al., 2002; Bonato et al., 2006; Wysmierski e Vello, 2013). Uma das consequências deste fato é a susceptibilidade da cultura a pragas, doenças e fatores ambientais que causam estresses bióticos e abióticos, ocasionando redução na produtividade da cultura.

Considerando as consequências da baixa variabilidade genética e com o objetivo de desenvolver genótipos mais produtivos e resistentes a pragas e doenças surgiu a necessidade de uso das ferramentas da genômica no estudo de genes de interesse por meio de marcadores moleculares de ampla cobertura no genoma, caso dos marcadores SNPs (do Inglês: Single Nucleotide Polymorphism). Os SNPs têm sido utilizados no estudo da diversidade genética, estrutura populacional, e principalmente de genes que controlam caracteres quantitativos, comumente conhecidos como QTL (do Inglês: Quantitative Trait Loci).

Uma das ferramentas utilizada nos últimos anos é o mapeamento associativo, baseado no uso do desequilíbrio de ligação (LD) dos genes com os SNPs, num conjunto de indivíduos de diferentes origens. Esta metodologia permite avaliar uma ampla cobertura do genoma, considerando os múltiplos eventos de

recombinação sucedidos ao longo da história evolutiva da espécie nas diversas populações, fato que permite associar genes de interesse aos marcadores SNPs.

Embora o mapeamento associativo tenha sido superior a outras metodologias, como o mapeamento de QTL usando populações segregantes, a avaliação de modelos de mapeamento por associação continua sendo um desafio à identificação de regiões genômicas de interesse e seu uso na seleção assistida por marcadores. O uso das informações de genótipos ausentes, de heterogeneidade genética, de desequilíbrio de ligação ou da baixa frequência alélica são fatores importantes para o desenvolvimento de modelos de associação estatisticamente apropriados para o estudo da arquitetura de caracteres complexos.

O uso das informações da estrutura populacional e do grau de relacionamento genético é comum no estudo de mapeamento por associação. Além disso, a informação de haplótipos baseados em SNPs constitui uma abordagem interessante pra o estudo de marcadores associados à caracteres agronômicos de interesse. O uso destes haplótipos nos modelos de associação é comumente mais efetivo em presença de desequilíbrio de ligação, dado à limitação bi-alélica dos marcadores SNPs. Consequentemente, o uso de haplótipos na análise de associação poderia melhorar a eficiência na detecção de associações significativas em caracteres agronômicos.

Assim, o objetivo do presente trabalho foi identificar regiões genômicas que controlam peso de 100-sementes, altura de planta e rendimento de grãos, em um painel de mapeamento associativo de soja, usando marcadores moleculares SNPs e informação de haplótipos.

2. REVISÃO DE LITERATURA

2.1. História e importância da cultura da soja

A soja silvestre (*Glycine soja* Sieb. and Zucc.) e a soja cultivada (*Glycine max* (L.) Merril) pertencem ao subgênero soja dentre do gênero *Glycine* L., sendo *G. soja* considerada o parente silvestre mais próximo de *G. max*, posto que ambos têm 20 cromossomos (2n = 40), hibridizam facilmente, exibem emparelhamento cromossômico meiótico normal e geram híbrido fértil viável (Kim et al., 2010). Acredita-se que a forma cultivada tenha derivado de *G. soja*, em razão do acúmulo de características qualitativas e quantitativas resultantes de mutações genéticas, sem alteração do número cromossômico (Bonetti, 1981). Esta informação tem sido corroborada pelos estudos morfológicos, citogenéticos e moleculares que indicam que a soja cultivada foi domesticada da soja silvestre na China (Broich e Palmer, 1980; Kollipara et al., 1997; Doebley et al., 2006).

Segundo Hymowitz (1970), a soja (*G. soja*) surgiu na região nordeste da China, por volta do século XVII a.C, e deste país expandiu-se para outras partes da Ásia, por volta do século XI a.C. Estudos complementares indicam que a distribuição geográfica limita-se para a Ásia Oriental, abrangendo vastas áreas da China, bem como as regiões adjacentes, incluindo o extremo Oriente russo, a Península da Coreia e do Japão (Singh e Hymowitz, 1988; Boerma e Specht, 2004).

G. max, inicialmente domesticada na China, em latitudes compreendidas entre 30 e 45ºN, foi posteriormente disseminada para a América do Norte, Europa e América do Sul. No Brasil, desde o final do século XIX e durante muitas décadas, foi plantada somente em caráter experimental por algumas instituições de pesquisa (Priolli et al., 2004). Já no século XX, a partir da década de 60, a cultura passou a adquirir importância no País, inicialmente na Região Sul (latitudes 30 a 22ºS), onde apresentou melhor adaptação, devido à semelhança com as regiões tradicionais de cultivo no mundo. A necessidade de aumento da produção e os bons resultados dos programas de melhoramento da cultura provocaram a rápida expansão da área de cultivo desta leguminosa da Região Sul rumo ao Cerrado, latitudes 20 a 5ºS (Urben Filho e Souza, 1993), levando o Brasil de uma posição inexpressiva no cenário mundial para a de segundo maior produtor de soja no mundo (Priolli et al., 2002).

A soja cultivada (*G. max*) é originária de climas temperados, mas nunca foi encontrada na forma silvestre (Borem, 2001). Mesmo assim, os diversos genótipos se adaptam bem em uma ampla faixa de outros climas, em razão da sua aclimatação aos climas tropical e subtropical. Isso faz supor que a soja apresente ampla diversidade genética quanto à sua área de adaptação e esta característica se deve principalmente à sensibilidade da cultura ao fotoperíodo e à temperatura (Sediyama e Santos, 1988).

A adaptação e desenvolvimento do cultivo da soja são influenciados pelas condições ambientais, por exemplo, temperatura, altitude, umidade relativa, condições pluviométricas, tipo de solo e fotoperíodo (Câmara, 1997). O fotoperíodo é o fator ambiental mais importante que interfere na passagem da soja do estádio vegetativo para o reprodutivo. Consequentemente, a sensibilidade da cultura ao fotoperíodo é um fator que dificulta o aumento da sua faixa de adaptação (Alliprandini et al., 2009).

A possibilidade de adaptação das diversas cultivares da soja para as regiões de cultivo no Brasil foi o foco principal nos programas de melhoramento do país. Neste contexto, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), considerando a diversidade de ecossistemas e tipos de solo e clima, apresentou uma proposta de regionalização dos testes de indicação de cultivares da soja para o Brasil, estabelecendo cinco macrorregiões sojícolas e vinte regiões edafoclimáticas para pesquisa e indicação de cultivares (Kaster e Farias, 2005; Alliprandini et al. 2009; Kaster e Farias, 2012).

No Brasil, o aumento da produção de soja foi observado a partir da década de 1970 e ocorreu devido à ação conjunta dos programas de fitotecnia e de melhoramento genético de várias instituições estabelecidas nas diversas regiões do País (Borem, 2001). Dessa maneira, o desenvolvimento de material genético apropriado para cada região tem sido um dos fatores responsáveis pelo progresso da produção na soja, a qual é cultivada em grande diversidade de ambientes, englobando altas e baixas latitudes.

A produtividade da soja tem aumentado significativamente nas últimas safras. Atualmente a soja representa 52% da produção total de grãos do Brasil. Segundo informação da CONAB, na safra 2016/2017, a soja apresentou uma estimativa de área semeada de 33,8 milhões de hectares, com produtividade média

de 3.480 kg ha^{-1}. Para a safra 2016/2017, a CONAB destacou a cultura da soja como a responsável pelo possível aumento de área semeada (CONAB, 2017).

2.2. Diversidade genética e coancestralidade

A diversidade genética de uma espécie é determinada pelas frequências alélicas observadas em nível de indivíduo, populações ou espécies. Esta medida permite determinar o nível de relacionamento genético, quantificar e prever o nível de variabilidade total existente e sua distribuição nas unidades taxonômicas avaliadas (sejam: indivíduos, acessos de bancos de germoplasma, linhagens, cultivares, populações ou espécies) (Weir, 1996).

O amplo uso do germoplasma nas distintas culturas, acompanhado do processo de domesticação, tem provocado uma redução na diversidade genética, trazendo a perda de genes úteis, reservados de parentes selvagens. Durante a domesticação, foram selecionadas as linhas que continham caracteres agronomicamente importantes, o que resultou na redução da diversidade alélica de todo o genoma (Tanksley e McCouch, 1997; Buckler et al., 2001).

Existem duas maneiras básicas de inferir sobre a diversidade genética, sendo a primeira de natureza quantitativa e a outra de natureza preditiva. Entre esses dois métodos de avaliação da diversidade, os métodos preditivos recebem maior atenção, pois dispensam a obtenção prévia das combinações híbridas (Cruz et al., 2011). Métodos preditivos têm sido utilizados em estudos de diversidade genética em soja, com objetivo de selecionar genótipos divergentes para a formação de híbridos superiores (Almeida et al., 2011; Rigon et al., 2012).

Metodologias multivariadas, tais como análise de componentes principais, variáveis canônicas e métodos de agrupamento, têm sido utilizadas amplamente no estudo da diversidade genética das distintas cultivares de soja, tanto no Brasil como em outros países (Priolli et al., 2004; Priolli et al., 2010). Estas metodologias são feitas com base em dados morfológicos e moleculares e segundo a natureza quantitativa ou qualitativa dos dados. Para o caso de variáveis contínuas, a análise é baseada nas medidas de dissimilaridade, tais como distância euclidiana, euclidiana média e de Mahalanobis, entre outras. No caso de variáveis categóricas, utilizam-se medidas de similaridade, como coeficiente de concordâncias simples e de Jaccard

(Cruz et al., 2011). Dentro dos métodos de agrupamento os mais utilizados até agora, tem sido o método hierárquico da ligação média entre grupos (UPGMA).

No Brasil, os primeiros trabalhos baseados em dados morfológicos estimaram que o germoplasma brasileiro da soja proviesse de base genética restrita, tendo se originado de poucas linhagens ancestrais. Bonetti (1981) estimou que cerca de 70% das cultivares desenvolvidas para o Rio Grande do Sul, na década de 1960, descendiam das cultivares americanas Hill, Hood ou ambas. Hiromoto e Vello (1986), utilizando o coeficiente de parentesco, informaram que todas as cultivares recomendadas para plantio naquele ano agrícola descendiam de 26 cultivares, sendo que, deste total, apenas quatro (CNS, Roanoke, Tokyo e S-100) eram responsáveis por cerca da metade daquele conjunto gênico (48,2%). Dessas 26 cultivares, seis (CNS, S-100, Roanoke, Tokyo, PI 54610 e Dunfield) foram compartilhadas com a base genética de América do Norte, segundo o estudo de Gizlice et al. (1994). Além disso, os cinco melhores ancestrais compartilhados entre ambos os países são as que dão maior contribuição para o germoplasma do sul dos EUA, que também é de base genética estreita (Gizlice et al., 1993; Gizlice et al., 1994).

Estudos baseados na distância genética estimada pelo uso de marcadores moleculares (Priolli et al., 2002; Bonato et al., 2006) e informação de pedigree (Miranda et al., 2007; Priolli et al., 2010) têm corroborado a informação dos dados morfológicos e também mostrado que uma coleção representativa de cultivares de soja recomendadas para o cultivo em todas as regiões brasileiras se agrupam de acordo com seu pedigree (Priolli et al., 2013). Segundo Wysmierski e Vello (2013), o principal ancestral, CNS, está presente no pedigree de 435 (98%) cultivares de soja no Brasil. Os outros ancestrais superiores (S-100, Roanoke e Tóquio) também têm frequências muito altas. CNS e S-100 são os ancestrais mais comuns porque do seu cruzamento resultou na cultivar LEE e na linha D49-2491, uma irmã de LEE e um antepassado de Bragg. LEE e Bragg foram utilizados como os principais genitores em muitas cultivares precoces desenvolvidas no Brasil.

Priolli et al. (2004), fazendo uso de marcadores moleculares microsatélites, avaliaram a variabilidade genética entre programas de melhoramento de soja no Brasil. Constataram que havia maior variabilidade genética dentro dos programas de melhoramento genético do que entre eles. Oliveira (2014) avaliou pela primeira vez o germoplasma da soja brasileira utilizando uma ampla cobertura do genoma com

marcadores SNPs e microssatélites, evidenciando que algumas cultivares de diferentes programas de melhoramento apresentaram uma similaridade maior que 95%, corroborando os resultados anteriores. Além disso, observou que cultivares adaptadas para a mesma região de cultivo estavam no mesmo grupo.

Além da existência de variabilidade genética reportada para as cultivares de soja convencionais estabelecidos no Brasil, Villela (2013) verificou a presença de variabilidade genética em amostra representativa das cultivares transgênicas (RR) cultivadas e comercializadas no país. Os resultados mostraram que, dos seis programas de melhoramento avaliados (Brasmax (7 cultivares avaliadas), Coodetec (5), Embrapa (19), Monsanto (12), Pionner (7) e TMG(11)), o da Monsanto apresentou maior diversidade genética. Os programas da Brasmax, Pionner e Coodetec apresentaram menor diversidade genética entre suas cultivares, mas o menor número de cultivares avaliadas dos programas Pionner, Coodetec e Brasmax pode ter colaborado para a menor diversidade apresentada nesses programas (Villela, 2013), consequentemente existe a necessidade de maiores estudos referidos a estimação da variabilidade nestes genótipos.

Na China, diversos estudos têm sido relatados a respeito da diversidade genética e a estrutura populacional da soja silvestre e cultivada (Xu et al., 2002; Lee et al., 2008; Li et al., 2009). Os estudos suportam a evidência de um único ou múltiplos eventos de domesticação, mas filogeneticamente a soja cultivada e silvestre estão representadas num único cluster, corroborando a origem monofilética de todas as cultivares da soja cultivada (*G. max*) (Guo et al., 2010). Esta informação poderia explicar em parte o porquê da estreita base genética das cultivares de soja compartilhadas entre o Brasil e EUA.

2.3. Estrutura genética e mapeamento associativo

A estrutura genética resulta da ação conjunta dos processos naturais de migração, mutação, seleção e deriva genética, que atuam dentro do contexto histórico e biológico de cada espécie (Falush et al., 2007). Em populações de soja, os métodos tradicionais de estimação da estrutura populacional estão baseados na comparação da diversidade genética de populações pré-definidas e de acordo a origem geográfica delas.

Diversos trabalhos têm estudado a estrutura populacional dos genótipos de soja. Os resultados sugerem que, durante o período glacial, uma expansão na rota entre o sudeste e nordeste da China poderia ter resultado em genótipos similares de soja selvagem nas populações originais daquelas regiões (Guo et al., 2012). Isso foi corroborado por Chung et al. (2013), cujos resultados da estrutura populacional baseada no re-sequenciamento de 10 cultivares e 6 cultivares silvestres da soja apoiam a hipótese de que a soja cultivada é um subclado de seu progenitor selvagem, mas também refuta a hipótese de múltiplos eventos de domesticação da soja na Ásia Oriental. Da mesma forma, Wang et al. (2013) observaram que não existe um ancestral intermediário entre a soja cultivada e a silvestre.

Apesar de a soja cultivada e a silvestre serem altamente relacionadas, as constantes hibridações têm gerado uma complexa estrutura populacional. Qiu et al. (2014) demostraram uma mistura de populações entre a soja cultivada, semi-silvestre (*Glycine gracilis*) e silvestre. Pelo contrário, uma estrutura populacional menos complexa tem sido encontrada em uma coleção de cultivares recomendadas para a cultura no Brasil. A análise bayesiana revelou a presença de dois clusters ou subgrupos, os quais foram agrupados de acordo com seus ancestrais de origem: Bragg, Hood e Santa Rosa, entre alguns dos pais do grupo 1; LEE, UFV1, Tropical e Paraná foram alguns dos pais do grupo 2 (Priolli et al., 2013). No entanto, estudos conduzidos com germoplasma de soja da China (Wang et al., 2006; Guo et al., 2012) e Estados Unidos (Ude et al., 2003) reconhecem que embora os clusters tenham ancestrais comuns, mas não foi possível identificar os ancestrais de origem para cada um dos clusters formados.

O efeito da estrutura populacional é o principal fator limitante nos estudos genéticos de mapeamento por associação (Pritchard et al., 2000), pois a presença de grupos geram falsas associações em nível populacional, devido à relação entre indivíduos da mesma espécie (Pritchard e Rosenberg, 1999). Ao invés de gerar associações genéticas genuínas com o caráter de interesse, as diferenças das frequências alélicas entre as subpopulações em uma população geram falsas associações. A fim de eliminar esse viés, diversos trabalhos feitos em soja, baseados na metodologia de análise bayesiana, permitiram determinar a relação da estrutura populacional entre indivíduos ou populações aparentadas (Guo et al., 2012; Priolli et al., 2013). Essa análise de correção da estrutura populacional é comum em estudos associativos que visem a identificar marcadores que contribuem

para o fenótipo (Hao et al., 2012b; Qiu et al., 2014). Neste caso, incorpora-se no modelo de associação o efeito da estrutura genética na forma de matriz de inclusão de cada indivíduo para um grupo especifico.

A não incorporação do efeito da estrutura populacional nos modelos de associação tem gerado associações falsas. Segundo Kassem et al. (2006) cerca do 85% dos QTLs reportados até uns anos atrás não puderam ser confirmados em estudos posteriores, e poucos têm sido realmente aplicados em programas de melhoramento. Segundo Wang et al. (2008) e Xu e Crouch (2008), isto ocorre porque a maioria dos QTLs foram estudados para populações específicas, e a variação genética detectada na população bi-parental específica não pode ser compartilhada em outras populações genéticas.

2.4. Desequilíbrio de ligação e mapeamento associativo

O mapeamento por associação é comumente conhecido como mapeamento por desequilíbrio de ligação por estar baseado na ligação de alelos específicos, marcadores moleculares ou haplótipos (combinação de genótipos em grupo de marcadores ligados) presentes em alta densidade no genoma em estudo, os quais podem ser associados com caracteres fenotípicos com alto nível de significância.

Historicamente, as análises de ligação fatorial têm sido comumente estudadas em populações controladas segregantes, altamente estruturadas e com pedigree conhecido (Lander e Botstein, 1989; Haley e Knott, 1992; Jansen et al., 1993; Zeng, 1993), mas este tipo de populações apresentam duas importantes limitações: o número limitado de eventos de recombinação estudado e o fato de que apenas dois alelos por loco podem ser avaliados simultaneamente. Ao contrário do caso anterior, o mapeamento associativo abriu a possibilidade de estudar o total de eventos de recombinação acontecidos ao longo da história evolutiva das populações de um germoplasma específico. A diferença entre o grau de possível detecção de QTL com estas populações naturais, em comparação ao uso da descendência de cruzamentos segregantes, é o nível de desequilíbrio de ligação (Hwang et al., 2014).

O desequilíbrio de ligação é a associação não casual de alelos em diferentes locos e a correlação entre polimorfismos que é causada pelos eventos de mutação e recombinação compartilhados entre as populações de uma espécie ao longo da sua história evolutiva. Não é necessário realizar análise de ligação, uma vez que, devido

à grande saturação do genoma com marcadores SNPs, assume-se que estes estão diretamente em LD com o QTL. Essa relação faz com que determinados marcadores moleculares amplamente distribuídos no genoma, assim como os marcadores SNPs, estejam associados com alelos particulares que afetam uma determinada característica de interesse.

Diferentes trabalhos têm considerado os marcadores moleculares SNPs para a associação com caracteres agronômicos de interesse em soja (Hao et al., 2012a; Hao et al., 2012b; Hwang et al., 2014). Isso ocorre, principalmente, por causa da ampla cobertura do genoma neste tipo de marcadores e pela sua natureza codominante, embora na soja cultivada apresente vários eventos de "gargalo genético" e dois grandes eventos de duplicação do genoma, que resultaram em menor diversidade das sequências genéticas, limitando a quantidade de SNPs (Schumtz et al., 2010). Mesmo assim, seu uso tem sido de ampla aplicabilidade no mapeamento associativo.

2.5. Uso de haplótipos no mapeamento associativo

Um bloco de haplótipo é uma região genômica na qual dois ou mais loci polimórficos ou marcadores SNPs, que apresentam estreita proximidade no genoma, tendem a ser herdados juntos (Abdel-Shafy et al., 2014). Segundo Greenspan e Geiger (2004), acredita-se que esses blocos sejam causados por *hotspots* de recombinação dentro de regiões do DNA, em que SNPs segregam consequentemente de uma geração para a seguinte, agindo como alelos combinados. A combinação de alelos de SNPs num bloco de haplótipo pode ter maior desequilíbrio de ligação com o alelo de um QTL do que os alelos de SNPs individuais utilizados para construir o haplótipo (Abdel-Shafy et al., 2014).

Lorenz et al. (2010) usaram dados fenotípicos simulados para mostrar que a metodologia de uso de haplótipos baseados em SNPs pode aumentar o poder do mapeamento por associação sobre o uso de marcadores de SNP individuais. De acordo com Song et al. (2015), espécies altamente endogâmicas, como a soja, são adequadas para o mapeamento por associação usando blocos de haplótipos. Neste caso, o uso de haplótipos na associação poderia melhorar a eficiência na detecção de associações significativas (Hao et al., 2012a; Hao et al., 2012b; Zhang et al., 2014).

2.6. Modelos para o mapeamento por associação

Diversas abordagens estatísticas têm sido utilizadas com o objetivo de identificar a associação entre marcadores moleculares e o caráter fenotípico de interesse avaliado em um grande número de indivíduos. Entre as metodologias mais importantes estão a análise de regressão linear simples e múltipla, análise de variância, o uso de modelos mistos e o teste t e qui-quadrado (Thornsberry et al., 2001; Yu et al., 2006; Zhu et al., 2008).

O grau de relacionamento genético que poderia existir entre os indivíduos que constituem a população de mapeamento representa o principal problema para a associação entre genótipo e fenótipo, pois gera falsas associações não funcionais (Pritchard et al., 2000; Thornsberry et al., 2001). Consequentemente, o uso da informação da estrutura populacional tem sido explorado nos modelos de associação (Thornsberry et al., 2001). Segundo Yu et al. (2006), idealmente, as populações que apresentam mínima estrutura populacional ou grau de parentesco, resultam em maior poder estatístico, desde que a característica de interesse esteja bem distribuída na população.

Na análise de mapeamento por associação, o uso de modelos mistos, representa uma abordagem estatística poderosa, que permite contornar e ou diminuir o efeito do grau de relacionamento genético (Yu et al., 2006). A primeira geração de modelos mistos aplicados na análise de mapeamento por associação utilizou, além da estrutura populacional (matriz Q), a informação de pedigree baseada em marcadores genéticos. Essa informação do grau de relacionamento genético gera uma matriz chamada de parentesco, ou simplesmente K (VanRaden, 2008).

Os modelos mistos têm sido usados há muito tempo em pesquisas da área genética (Henderson, 1984) e, mais especificamente, os métodos de mapeamento por associação de modelo misto foram desenvolvidos pra a dissecção de rasgos complexos em diferentes espécies (Zhao et al., 2007; Zhu et al., 2008). Neste contexto, as estatísticas da análise de deviance e o critério de informação bayesiano (do Inglês: *Bayesian Information Criteria* ou BIC) (Schwarz, 1978) têm sido propostos como critérios para a seleção de modelos mistos no mapeamento por associação (Broman e Speed, 2002; Littell et al., 2006).

Os critérios de informação foram desenvolvidos com base na teoria da informação; um ramo da matemática aplicada, relacionada à quantificação (o processo de contagem e medição) de informações. Para o caso da seleção de modelos mistos de mapeamento por associação, o modelo selecionado será aquele que minimize o valor de determinado critério de informação (BIC ou outro). Um modelo completo (com todos os efeitos) será escolhido se tiver o menor valor do critério de informação comparado com o valor do modelo reduzido (modelo descontando um determinado efeito).

3. REFERENCIAS BIBLIOGRÁFICAS

ABDEL-SHAFY, H.; BORTFELDT, R.H.; TETENS, J.; BROCKMANN, G.A. Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. **Genetics Selection Evolution,** 46:35, 2014.

ALLIPRANDINI, L.F.; ABATTI, C.; BERTAGNOLLI, P.F.; CAVASSIM, J.E.; GABE, H.L.; KUREK, A.; MATSUMOTO, M.N.; DE OLIVEIRA, M.A.R.; PITOL, C.; PRADO, L.C.; STECKLING, C. Understanding soybean maturity groups in Brazil: environment, cultivar classification, and stability. **Crop Science**, 49:801–808, 2009.

ALMEIDA, R.D.; PELUZIO, J.M.; AFFÉRRI, F.S. Divergência genética entre cultivares de soja, sob condições de várzea irrigada, no sul do Estado Tocantins. **Revista Ciência Agronômica**, 42:108-115, 2011.

BOERMA, H.R.; SPECHT, J.E. **Soybeans: Improvements, production and uses**. Wisconsin: Madison, 2004. 1144p.

BONATO, A.L.V.; CALVO, E.S.; GERALDI, I.O.; ARIAS, C.A.A. Genetic similarity among soybean (*Glycine max* (L) Merrill) cultivars released in Brazil using AFLP markers. **Genetics and Molecular Biology,** 29:692-704, 2006

BONETTI, L.P. Distribuição da Soja no Mundo. In: MIYASAKA, S.; MEDINA J.C. (eds.). **A soja no Brasil**. Campinas: ITAL, 1981. cap. 1, Item 1, p.1-6.

BOREM, A. **Melhoramento de Plantas**. Viçosa: Editora UFV, 2001. 500p.

BROICH, S.L.; PALMER, R.G. A cluster analysis of wild and domesticated soybean phenotypes. **Euphytica,** 29:23-32, 1980.

BROMAN, K.W.; SPEED, T.R. A model selection approach for the identification of quantitative trait loci in experimental crosses. **Journal of the Royal Statistical Society Series B,** 64:641–656, 2002.

BUCKLER, E.S.; THORNSBERRY, J.M.; KRESOVICH, S. Molecular diversity, structure and domestication of grasses. **Genetics Research,** 77:213-218, 2001.

CÂMARA G.M.S. **Efeito do fotoperíodo e da temperatura no crescimento, florescimento e maduração de cultivares de soja (*Glycine max* (L.) Merrill)**. Viçosa: Universidade Federal de Viçosa, 1991. 266p. Tese (Doutorado em Fitotecnia).

CHUNG, W.H.; JEONG, N.; KIM, J.; LEE, W.K.; LEE, Y.G.; LEE, S.H.; YOON, W.; KIM, J.H.; CHOI, I.Y.; CHOI, H.K.; MOON, J.K.; KIM, N.; JEONG, S.C. Population Structure and Domestication Revealed by High-Depth Resequencing of Korean Cultivated and Wild Soybean Genomes. **DNA Research,** 21:153-167, 2013.

CONAB. Companhia Nacional de Abastecimento. **Acompanhamento da safra brasileira: grãos safra 2015/2016, sexto levantamento, março/2017**. Brasília: Conab, 2016. 25p.

CRUZ, C.D.; FERREIRA, F.M.; PESSONI, L.A. **Biometria aplicada ao estudo da diversidade genética**. Visconde do Rio Branco: Suprema, 2011. 620p.

DOEBLEY, J.F.; GAUT, B.S.; SMITH, B.D. The molecular genetics of crop domestication. **Cell,** 127:1309–1321, 2006.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. **Molecular Ecology Notes,** 7:574-578, 2007.

GIZLICE, Z.; CARTER, T.E.; BURTON, J.W. Genetic base for North American public soybean cultivars released between 1947 and 1988. **Crop Science,** 34:1143–1151, 1994.

GIZLICE, Z.; CARTER, T.E.; BURTON, J.W. Genetic diversity in North American soybean: I. Multivariate analysis of founding stock and relation to coefficient of parentage. **Crop Science,** 33:614–620, 1993.

GREENSPAN, G.; GEIGER, D. Model-based inference of haplotype block variation. **Journal of Computational Biology,** 11:493-504, 2004.

GUO, J.; FEI, L.Y.; WANG, Y.; CHEN, J.; LI, Y.; HUANG, H.; QIU, L.; WANG, Y. Population structure of the wild soybean (*Glycine soja*) in China: implications from microsatellite analyses. **Annals of Botany,** 110:777 –785, 2012.

GUO, J.; WANG, Y.; SONG, C.; ZHOU, J.; QIU, L.; HUANG, H.; WANG, Y. A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. **Annals of Botany,** 106:505–514, 2010.

HALEY, C.S.; KNOTT, S.A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. **Heredity,** 69:315-324, 1992.

HAO, C.; WANG, Y.; HOU, J.; FEUILLET, C.; BALFOURIER, F.; ZHANG, X. Association mapping and haplotype analysis of a 3.1-Mb genomic region involved in Fusarium head blight resistance on wheat chromosome 3BS. **PLoS ONE,** 7:e46444, 2012a.

HAO, D.; CHENG, H.; YIN, Z.; CUI, S.; ZHANG, D.; WANG, H.; YU, D. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. **Theoretical and Applied Genetics,** 124:447-458, 2012b.

HIROMOTO, D.M.; VELLO, N.A. The genetic base of Brazilian soybean cultivars. **Brazilian Journal Genetic,** 9:295-306, 1986.

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

HYMOWITZ, T. On the domestication of soybean. **Economic Botany,** 24:421-480, 1970.

JANSEN, R.C. Interval mapping of multiple quantitative trait loci. **Genetics,** 135:205-211, 1993.

KASSEM, M.; SHULTZ, J.; MEKSEM, K.; CHO, Y.; WOOD, A.; IQBAL, M.; LIGHTFOOT, D. An updated ‘Essex’ by ‘Forrest’ linkage map and first composite

interval map of QTL underlying six soybean traits. **Theoretical and Applied Genetics,** 113:1015-1026, 2006.

KASTER, M.; FARIAS, J.R.B. **Regionalização dos testes de valor de cultivo e uso e da indicação de cultivares de soja – segunda aproximação.** Londrina: Embrapa Soja, 2005. 12p.

KASTER, M.; FARIAS, J.R.B. **Regionalização dos testes de valor de cultivo e uso e da indicação de cultivares de soja – terceira aproximação**. Londrina: Embrapa Soja, 2012. 69p.

KIM, M.Y; LEE, S.; VAN, K.; KIM, T.H.; JEONG, S.C.; CHOI, I.Y.; KIM, D.S.; LEE, Y.S. PARK, D.; MA, J.; KIM, W.Y.; KIM, B.C.; PARK, S.; LEE, K.A.; KIM, D.H.; KIM, K.H.; SHIN, J.H.; JANG, Y.E.; KIM, K.D.; LIU, W.X.; CHAISAN, T.; KANG, Y.J.; LEE, Y.H.; KIM, K.H.; MOON, J.K.; SCHMUTZ, J.; JACKSON, S.A.; BHAK, J.; LEE, S.H. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. **Proceeding of the National Academy of Science USA,** 107:22032–22037, 2010.

KOLLIPARA, K.P.; SINGH, R.J.; HYMOWITZ, T. Phylogenetic and genomic relationships in the genus *Glycine* Wild. based on sequences from the ITS region of nuclear rDNA. **Genome,** 40:57-68, 1997.

LANDER, E.S.; BOTSTEIN, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics,** 121:185-199, 1989.

LEE, J.D.; YU, J.K.; HWANG, Y.H.; BLAKE, S.; SO, Y.S.; LEE, G.J.; NGUYEN, H.T.; SHANNON, J.G. Genetic diversity of wild soybean (*Glycine soja* Sieb. and Zucc.) accessions from South Korea and other countries. **Crop Science,** 48:606-616, 2008.

LI, X.H.; WANG, K.J.; JIA, J.Z. Genetic diversity and differentiation of Chinese wild soybean germplasm (*G. soja* Sieb. & Zucc.) in geographical scale revealed by SSR markers. **Plant Breeding,** 128:658-664, 2009.

LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER, R.D.; SCHABENBERGER, O. **SAS for Mixed Models**. SAS Press: Cary, NC, USA, 2006.

LORENZ, A.J.; HAMBLIN, M.T.; JANNINK, J.L. Performance of single nucleotide polymorphisms versus haplotypes for Genome-Wide Association analysis in Barley. **PLoS ONE,** 5:e14079, 2010.

MIRANDA, Z.F.S.; ARIAS, C.A.A.; PRETE, C.E.C.; KIIHL, R.A.S.; ALMEIDA, L.A.; TOLEDO, J.F.F.; DESTRO, D. Genetic characterization of ninety elite soybean cultivars using coefficient of parentage. **Pesquisa Agropecuária Brasileira,** 42:363-369, 2007.

OLIVEIRA, M. **Análise de desequilíbrio de ligação e diversidade genética em soja utilizando marcadores moleculares**. Maringá: Universidade Estadual de Maringá, 2014. 61p. Tese (Doutorado em Genética e Melhoramento).

PRIOLLI, R.H.G.; MENDES-JUNIOR, C.T.; ARANTES, N.E.; CONTEL, E.P.B. Characterization of Brazilian soybean cultivars using microsatellite markers. **Genetics and Molecular Biology,** 25:185-193, 2002.

PRIOLLI, R.H.G.; MENDES-JUNIOR, C.T.; SOUSA, S.M.B.; SOUSA, N.E.A.; CONTEL, E.B.P. Diversidade genética da soja entre períodos e entre programas de melhoramento no Brasil. **Pesquisa Agropecuária Brasileira,** 39:967-975, 2004.

PRIOLLI, R.H.G.; PINHEIRO, J.B.; ZUCCHI, M.I.; BAJAY, M.M.; VELLO, N.A. Genetic diversity among Brazilian soybean cultivars based on SSR loci and pedigree data. **Brazilian Archives of Biology and Technology** 53:519-531, 2010.

PRIOLLI, R.H.G.; WYSMIERSKI, P.T.; DA CUNHA PINTO, C.; PINHEIRO, J.B.; VELLO, N.A. Genetic structure and a selected core set of Brazilian soybean cultivars. **Genetics and Molecular Biology,** 36:382-390, 2013.

PRITCHARD, J.K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics,** 155:945-959, 2000.

QIU, J.; WANG, Y.; WU, S.; WANG, Y.Y.; YE, C.Y.; BAI, X.; LI, Z.; YAN, C.; WANG, W.; WANG, Z.; SHU, Q.; XIE, J.; LEE, S.H.; FAN, L. Genome re-sequencing of semi-wild soybean reveals a complex soja population structure and deep introgression. **PLoS ONE,** 9:e108479, 2014.

RIGON, J.P.G.; CAPUANI, S.; BRITO NETO, J.F.; ROSA, G.M.; WASTOWSKI, A.D.; RIGON, C.A.G. Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. **Revista Ceres,** 59:233-240, 2012.

SCHUMTZ, J.; CANNON, S.; SCHLUETER, J.; MA, J.; MITROS, T.; NELSON, W.; HYTEN, D.; SONG, Q.; THELEN, J.; CHENG, J. Genome sequence of the palaeopolyploid soybean. **Nature,** 463:178–183, 2010.

SCHWARZ, G. Estimating the dimension of a model. **Annals Statistics,** 6:461-464, 1978.

SEDIYAMA, T. **Melhoramento de plantas**. Londrina: Editora Mecenas, 2015. 352p.

SEDIYAMA, T.; SANTOS, O.S. Escolha de cultivares. In: SANTOS O.S. (ed.). **A cultura da soja**. Rio de Janeiro: Editora Globo, 1988. p. 93-108.

SINGH, R.J.; HYMOWITZ, T. The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. **Theoretical and Applied Genetics,** 76:705-711, 1988.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Fingerprinting soybean germplasm and its utility in genomic research. **Genes Genomes Genetics,** 5:1999-2006, 2015.

TANKSLEY, S.D.; McCOUCH, S.R. Seed banks and molecular maps: unlocking genetic potential from the wild. **Science,** 277:1063– 1666, 1997.

THORNSBERRY, J.M.; GOODMAN, M.M.; DOEBLEY, J.; KRESOVICH, S.; NIELSEN, D.; BUCKLER, E.S. Dwarf8 polymorphisms associate with variation in flowering time. **Nature Genetics**, 28:286–289, 2001.

UDE, G.N.; KENWORTHY, W.J.; COSTA, J.M.; CREGAN, P.B.; ALVERNAZ, J. Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. **Crop Science,** 43:858-1867, 2003.

URBEN FILHO, G.; SOUZA, P.I.M. Manejo da cultura da soja sob cerrado: época, densidade e profundidade de semeadura. In: ARANTES, N.E.; SOUZA, P.I.M. (eds.). **Cultura da soja nos cerrados**. Piracicaba: Potafós, 1993, p. 267-298.

VANRADEN, P.M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91:4414-4423, 2008.

VILLELA, O.T. **Diversidade fenotípica e molecular de cultivares brasileiras de soja portadoras de gene RR**. São Paulo: Universidade Estadual Paulista, 2013. 80p. Dissertação (Mestrado em Agronomia).

WANG, J. Coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. **Molecular Ecology Resource,** 11:141–145, 2010.

WANG, J.; McCLEAN, P.; LEE, R.; GOOS, R.; HELMS, T. Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. **Theoretical and Applied Genetics,** 116:777–787, 2008.

WANG, K.J.; LI, X.H. Genetic diversity and gene flow dynamics revealed in the rare mixed populations of wild soybean (*Glycine soja*) and semi-wild type (*Glycine gracilis*) in China. **Genetic Resource Crop Evolution,** 60:2303-2318, 2013.

WANG, L.X.; GUAN, R.X.; LIU, Z.X.; CHANG, R.Z.; QIU, L.J. Genetic diversity of chinese cultivated soybean revealed by SSR markers. **Crop Science,** 46:1032-1038, 2006.

WEIR, B.S. **Genetic data analysis II**. Sunderland, MA: Sinauer Associates. 1996.

WYSMIERSKI, P.T.; VELLO, N.A. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. **Genetics and Molecular Biology,** 36:547-555, 2013.

XU, D.X.; ABE, J.A.; GAI, J.G.; SHIMAMOTO, Y.S. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. **Theoretical and Applied Genetics,** 105:645–653, 2002.

XU, Y.; CROUCH, J. Marker-assisted selection in plant breeding: from publications to practice. **Crop Science,** 48:391-407, 2008.

YU, J.; PRESSOIR, G.; BRIGGS, W.; VROH, B.I.; YAMASAKI, M.; DOEBLEY, J.; MCMULLEN, M.; GAUT, B.; NIELSEN, D.; HOLLAND, J.; KRESOVICH, S.; BUCKLER, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics,** 38:203-208, 2006.

ZENG, Z.B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. **Proceeding of the National Academy of Science USA,** 90:10972-10976, 1993.

ZHANG, D.; KAN, G.; HU, Z.; CHENG, H.; ZHANG, Y.; WANG, Q.; WANG, H.; YANG, Y.; LI, H.; HAO, D.; YU, D. Use of single nucleotide polymorphisms and haplotypes to identify genomic regions associated with protein content and water-soluble protein content in soybean. **Theoretical and Applied Genetics,** 127:1905-1915, 2014.

ZHAO, K.; ARANZANA, M.J.; KIM, S.; LISTER, C.; SHINDO, C.; TANG, C.; TOOMAJIAN, C.; ZHENG, H.; DEAN, C.; MARJORAM, P.; NORDBORG, M. An Arabidopsis example of association mapping in structured samples. **PLoS Genetics,** 3:e4, 2007.

ZHU, C.; GORE, M.; BUCKLER, E.S.; YU, J. Status and Prospects of Association mapping in Plants. **The Plant Genome**, 1: 5-20, 2008.

CHAPTER 1 - POPULATION STRUCTURE, GENETIC RELATEDNESS AND LINKAGE DISEQUILIBRIUM BLOCKS IN TROPICAL SOYBEAN CULTIVARS (*GLYCINE MAX*)

ABSTRACT

Soybean (*Glycine max* L.) is an annual, self-pollinated species, whose genetic base in Brazil is result of several cycles of selection and effective recombination among a relatively small number of genotypes selected from the USA cultivars. This frequent selection, admixed population and the crossing of a small number of cultivars can increase the genetic relationship and affect the patterns of population structure. These factors affect the patterns of linkage disequilibrium (LD) blocks, which can be an effective approach for the screening of target loci for agricultural traits in cultivars of tropical soybean. The objectives of this research were to analyze LD blocks, estimate population structure and relationship through of genotyping of 169 cultivars of tropical soybean by using a BARCSoy6K of Illumina iScan platform. The genotyping revealed a high genetic relationship and population structure among the cultivars of soybean in Brazil, suggesting the existence of a shared genetic base. Our results provide a help to understand the distribution of genetic variation contained within the Brazilian soybean cultivar collection. The extensive use of a small number of elite genotypes in Brazilian breeding program further reduced the genetic variability, generate extensive LD and probably increase the haplotype block size. These results are in agreement with results of USDA soybean collection, mainly with American accessions when compared with wild and landraces accessions. We constructed a small haplotype block maps (941 bocks), which may be useful for association studies aimed at the identification of genes controlling traits of economic importance in soybean.

Keywords: Haplotypes; SNP; coancestry; self-pollination; soybean germplasm.

1. INTRODUCTION

Soybean (*Glycine max* L.) is an annual, self-pollinated species with a genome size of 1,115 Mpb (Schmutz et al., 2010). The species is believed to have originated from wild soybean *Glycine soja*, considering that both have 20 chromosomes (2n = 40), hybridize easily, exhibit normal meiotic chromosome pairing and generate viable fertile hybrids (Kim et al., 2010). The exact region of origin of soybean is still unknown, but southern China, the Yellow River valley of central China, northeastern China and several other regions are all candidate sources because *G. soja* grows naturally in far eastern Russia, China, Korea and Japan (Carter et al., 2004).

G. max is generally considered to have been domesticated from its wild relative (*G. soja*) 6,000~9,000 years ago in China (Carter et al., 2004) and may have been introduced to Korea, and then to Japan approximately 2,000 years ago, to North America in 1765, and to Central and South America during the first half of the last century. In this process of domestication and selection, a severe genetic bottleneck during soybean domestication was also found in several independent analyses (Xu et al., 2002; Hyten et al., 2006). There is supporting evidence for both single and multiple domestication events (Hymowitz and Kaizuma, 1981; Gai et al., 2000; Xu and Gai, 2003), which has been accompanied by a reduction in genetic diversity, as well as loss of useful traits reserved in wild relatives. This reduction of genetic diversity is common in crops have been subjected to strong selective pressure directed at genes controlling traits of agronomic importance during their domestication and subsequent episodes of selective breeding (i.e: Maize-Vigouroux et al., 2002).

The largest resource of soybean germplasm is the Asian landraces of *G. max* that are the most immediate result of domestication (Hyten et al., 2007). Selection, hybridization and breeding from these landraces have resulted in the release of improved cultivars in North American-USA (Gizlice et al., 1994). These first cultivars developed in USA were introduced and planted in Brazil during the 1960s and 1970s. With the growing importance of soybean, breeders began crossing these cultivars among themselves and with other sources, obtaining the first Brazilian cultivars, such as Industrial, Santa Rosa and Campos Gerais (Hiromoto and Vello, 1986). Thus, the

current Brazilian soybean germplasm pool, as defined by Hiromoto and Vello (1986), is the result of several cycles of selection and effective recombination among a relatively small number of selections from the USA cultivars.

The frequent selection, admixed population, and the crossing of a small number of cultivars in the Brazilian soybean breeding programs can lead to a reduction in genetic diversity and affect the patterns of linkage disequilibrium (LD). At the moment, few genetic studies have determined the patterns of LD in tropical soybean genotypes. Priolli et al. (2014), using 142 SSR markers and 94 accessions (cultivated and breeding material) obtained of EMBRAPA soybean and USP/ESALQ germplasm that represent soybean breeding lines of public and private institutions, suggest a structure of LD across the soybean genome (LD decay) of approximately 12 cM. In self-pollinated species, as well as soybean, where recombination is less effective than in outcrossing species, LD declines more slowly (Flint-Garcia et al., 2003). Nonetheless, the germplasm that makes up the collection plays a key role in LD variation because the extent of LD is influenced by the level of genetic variation captured by the target population (Soto-Cerda et al., 2013). In soybean, highly variable patterns of LD has been reported in multiple populations, with variability at different genomic regions (Hyten et al., 2007). In fact, due to the highly variable levels of LD decay in the Landraces and the Elite Cultivars reported for soybean (Hyten et al., 2007; Zhou et al., 2015) and the demands of dense marker sets, it is necessary to determine the LD in tropical soybean cultivars of Brazil that represent the range of photoperiod/temperature latitudinal adaptation as defined by a maturity group (MG) Roman numeral designation.

Most of the process observed in population genetics, as well as domestication, selection, founding events and population subdivision can affect LD decay, however, population structure (admixture) and the mating system of the species (selfing versus outcrossing) can strongly influence patterns of LD (Flint-García et al., 2003). It is known that pairwise LD increases with selfing and can extend very far in highly selfed organisms (Nordborg, 2000). For this reason, assume that individuals in a sample are either fully outcrossing may result in spurious inference of population structure in partially selfing populations, as suggested by Falush et al. (2003). To correct spurious evidence for admixture in the presence of partial self-fertilization, Gao et al. (2007) implement a model to accommodate partial selfing and correct the inference of population structure in self-pollinating species as

soybean. On the other hand, predict LD decay based on the present-day mating system must be cautious, because the mating system may have changed significantly (Flint-García et al., 2003). For example, *G. max* and its ancestor, *G. soja*, differ significantly in their outcrossing rates. The self-pollinating *G. max* has an outcrossing rate of approximately 1%, whereas *G. soja* outcrosses at an average rate of 13% (Fujita et al., 1997). The greater amount of outcrossing in *G. soja* increases the effective recombination rate, leading to the prediction of an 11-fold lower extent of LD in *G. soja* as compared to *G. max* (Flint-García et al., 2003).

In this study we genotyped 169 tropical soybean genotypes using high throughput genotyping with SNPs markers. The overall goal was to analyze linkage disequilibrium blocks in a collection of tropical soybean genotypes of Brazil. Our specific goals were: (1) to estimate the levels of population structure and assess population relatedness; (2) and to detect the patterns of LD blocks.

2. MATERIAL AND METHODS

2.1. Plant material and DNA extraction

A total of 169 cultivars of soybean with commercial use in Brazil were used for genotyping (Table S1). These cultivars represent the core cultivars used for Brazilian farmers from 1990s to 2010s, and some of these were important genitors in soybean breeding program of Brazil. Additionally, these cultivars were chosen to represent a range of materials developed for the Brazilian production area and representing the range of photoperiod/temperature latitudinal adaptation as defined by a maturity group (MG) Roman numeral designation (Table S1).

2.2. SNPs genotyping

Genomic DNA was extracted from leaf tissues collected from a mix of ten plants of each accession. DNA-easy Plant Kit (Qiagene) was used to DNA extraction. A total of 6,000 single nucleotide polymorphism (SNP) was genotyped in the 169 cultivars with an Infinium iSelect HD Custom Genotyping BARCSoy6K (Illumina Inc., San Diego, CA, USA) on the Illumina iScan platform. Genotyping was conducted by Deoxi Biotechnology Ltda ®, in Araçatuba, Sao Paulo, Brazil. After eliminating: redundant, non-polymorphic SNPs and SNPs with heterozygous alleles considered as missing data, a total of 4,949 SNPs remained. In addition, markers with MAF < 0.1 were removed from the genotype data set, leaving 3,780 SNPs for the population structure, coancestry and LD analysis.

2.3. Linkage disequilibrium

Linkage disequilibrium parameter (r^2) for estimating the degree of LD between pair-wise SNPs was calculated using the software TASSEL4.0 for each chromosomal and LD decay graph was plotted with physical distance (Mbp) vs r^2 for all intra-chromosomal comparison using nonlinear regression as described by Remington et al. (2001). The expected value of r^2 was estimate according to the following equation:

$$E(r^2) = \left[\frac{10+C}{(2+C)(11+C)}\right]\left[1+\frac{(3+C)(12+12C+C^2)}{n(2+C)(11+C)}\right]$$

Where r^2 = squared correlation coefficient, n = sample size, and C is a model coefficient for the distance variable (Hill and Weir, 1988). The LD decay curve was fitted to predicted r^2 values between adjacent markers using the model of Hill and Weir (1988). This model was implemented to determine LD decay as a function of the distance using the 'nlm' function in R. To determine the baseline r^2 values, a critical value of LD decay was calculated to 50 % of its initial value according to Mamidi et al. (2011) and Wen et al. (2015).

2.4. Linkage disequilibrium blocks analysis

The pairwise estimates D' and r^2 were calculated by chromosome. LD blocks were estimated by Solid Spine of LD using the software Haploview 4.2 (Barrett et al., 2005). This internally developed method of Haploview searches for a "spine" of strong LD running from one marker to another along the legs of the triangle in the LD chart. A cutoff of 1% was used, meaning that if addition of a SNP to a block resulted in a recombinant allele at a frequency exceeding 1%, then that SNP was not included in the block.

2.5. Population structure

Population structure and inbreeding coefficients at population level were estimated under the Markov Chain Monte Carlo (MCMC) algorithm for the generalized Bayesian clustering method implemented in InStruct software (Gao et al., 2007). This method does not assume Hardy-Weinberg equilibrium within loci, and the expected genotype frequencies are estimated based on rates of inbreeding or selfing.

For infer population structure and population selfing rates in soybean, we performed the function (mode) two of InStruct software (Gao et al., 2007). In fact, we implemented one independent run of MCMC sampling for numbers of groups (K parameter) varying from 2 to 10, without prior population information, and burn-in of 5,000 with run length periods of 50,000 iterations. The best estimate of number of

K groups was determined according to the lowest value of Deviance Information Criterion (DIC) among the nine K simulated (Gao et al., 2007). The hierarchical F statistics were used to estimate proportion of genetic variance explained by MG class and company of origin of soybean using ancestry estimates for K=9 and calculated using the hierfstat R package (Goudet, 2005).

2.6. Molecular coancestry

Strong relatedness among familial, subpopulations and populations can potentially cause spurious association when it is not considered in association mapping model. Relatedness between subpopulations was estimated using Reynolds genetic distance (Θ), which is given by Θ_{ij} =-ln (1-Fst) for subpopulations i and j (Reynolds et al., 1983), where F_{st} correspond to genetic differentiation among subpopulations. Pairwise molecular coancestry between the nine subpopulations of tropical soybean obtained previously with InStruct software was performed in the software Arlequin 3.5 (Excoffier and Lischer, 2010) using a total of 3,780 SNPs markers.

3. RESULTS AND DISCUSSION

3.1. Tropical soybean genotyping

A moderate coverage of the tropical soybean genome was obtained with the BARCSoy6K. In mean 247.5 SNPs markers were found by chromosome, with variation from 198 (chromosome 1) to 323 (chromosome 8). For each chromosome was estimated the ratio between the number of SNPs and the length of each chromosome measured in cM. On average, was found one SNP marker every 0.48 cM, ranging from 0.33 cM (chromosome 4) to 0.60 cM (chromosome 17) by SNP (Table 1). The most marker coverage was found for chromosome 8 that had 323 markers with an average marker density of 0.49 cM. In contrast, the chromosome 1 had the least number of SNPs markers equal to 198, with an average marker density of 0.55 cM. This demonstrates that Illumina Infinium platform of genotyping identified SNPs that were well distributed throughout the tropical soybean genome.

Table 1 - Distribution of SNPs markers and linkage disequilibrium blocks in the 20 chromosomes (Chr) in cultivars of tropical soybean

Chr	LG[†]	Length (cM) [†]	Number of SNPs	Average marker density cM/SNP	Blocks in Disequilibrium[‡]	Number of SNPs in LD Blocks
1	D1a	109.32	198	0.552	32	109
2	D1b	143.61	293	0.490	66	214
3	N	106.12	216	0.491	47	154
4	C1	75.06	225	0.334	41	135
5	A1	96.47	237	0.407	40	141
6	C2	147.50	258	0.572	54	167
7	M	134.00	262	0.511	49	174
8	A2	156.88	323	0.486	52	202
9	K	102.96	219	0.470	38	127
10	O	139.36	235	0.593	44	143
11	B1	135.09	227	0.595	42	133
12	H	120.18	205	0.586	36	132
13	F	144.67	313	0.462	65	206
14	B2	106.11	235	0.452	48	143
15	E	104.43	264	0.396	49	156
16	J	91.62	209	0.438	42	133
17	D2	128.40	218	0.589	35	115
18	G	110.91	321	0.346	74	223
19	L	111.59	260	0.429	50	155
20	I	124.34	231	0.538	37	124
TOTAL		2,388.613	4,949	0.487	941	3,086

[†]Source: Soybase (www.soybase.org).

[‡]Based on 3,780 SNPs markers.

Our SNP genotyping for tropical soybean is the first using Infinium BARCSoy6K. At the moment, this assay is being applied to genotype 169 accessions of tropical soybean that represent the whole germplasm of Brazil. This resulting dataset demonstrate the moderate coverage of tropical soybean genome by using this 6k SNP assay and will assist in the application of genome-wide association studies and high-resolution genetic linkage maps of important traits. Currently, two studies showed the utility of Illumina Infinium BARCSoy6K to understand the genetic architecture of complex traits and for identify SNPs tightly linked to quantitative trait loci (QTL) in many important soybean traits (Akond et al., 2013; Lee et al., 2015).

Some loci were found in heterozygosity. The percentage of heterozygosity was ranging from 3% (BRSMT Crixás, CD 205, P98Y70 and Celeste) to 41% (BMX Titan RR), with a mean of 9% among the 169 cultivars. Seventy-six percent of the cultivars (129) had fewer than 10% of heterozygosity; twenty percent (33) was between 10 to 30% and four percent had more than 35% of heterozygosity (Figure 1).

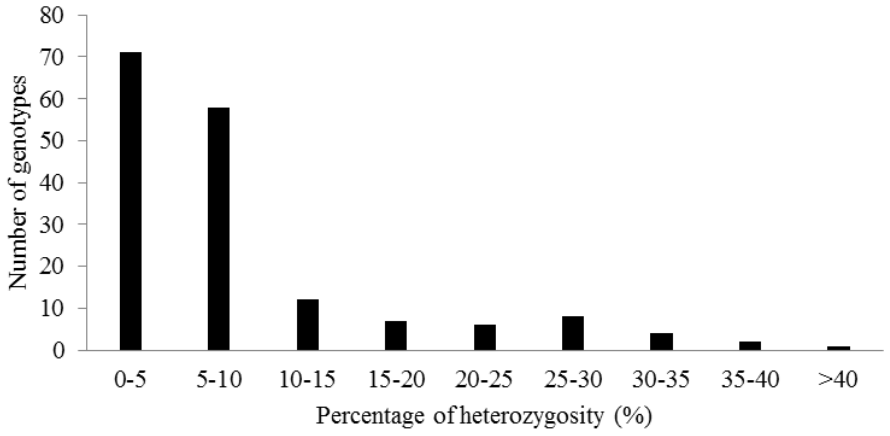


Figure 1 - Frequency of observed heterozygosity in 169 cultivars of soybean, using 4,949 SNPs markers.

In average, we found 9% of heterozygosity among the cultivars, which may be considered moderately high respect to other studies in soybean (Hyten et al., 2010), in addition it represent an important source of genetic diversity and adaptive evolution. Genetic theory predicts, on average, a halving of heterozygous loci with every self-pollination following a given cross. However, heterozygosity may be retained at higher rates if loci confer desirable and selectable phenotypes (Gore et

al., 2009), as the case of continuous selections of soybean cultivars in Brazil for different traits. Our result reveals high levels of heterozygosity in some cultivars of tropical soybean and it may be useful to promote genetic variability among the genetic base of soybean in Brazil. In fact, a recent study using phenotypic and molecular data (SSR markers) verified the existence of genetic variability among RR soybean cultivars in public and private soybean breeding companies of Brazil (Villela et al., 2014).

3.2. Population structure and molecular coancestry

The genetic structure of the 169 tropical soybean cultivars was estimated using a bayesian clustering approach to infer the number of strongly differentiated genetic subpopulations. According to DIC value, our population structure analysis supported the existence of nine subpopulations that come from different genetic breeding programs of Brazil (Figure 2, Table S1). Each subpopulation (K=9) contained admixed cultivars that come from different soybean genetic breeding programs of Brazil (Table 2). Nearly half of these were considered admixed because the degree of membership within a subpopulation was <0.5. Although 169 cultivars were used in this study, we were only able to obtain the pedigrees for 89 cultivars (Table S1), due to the Variety Protection Act of 1997, many breeders have not made public the pedigrees of released cultivars, especially more recent ones. However, our result reveals the existence of shared genetic base among the public and private breeding programs of soybean in Brazil, and showed the high genetic relationship that exist among the commercial cultivars.

A previous study conducted by Hiramoto and Vello (1996) indicate that Brazilian soybean ancestors have a narrow genetic base, with only four ancestors (CNS, S-100, Roanoke and Tokyo), that represent approximately 48% of the overall genetic base. Wysmierski and Vello (2013), evaluating 444 cultivars available in the database for the National Cultivar Registry from the Ministry of Agriculture, Livestock and Food Supply of Brazil, showed an increasing in the number of ancestors over time (1971 to 2009); however the same four main ancestors contribute more than half (55.3%) to the genetic base in soybean and were the same over 1971 to 2009, showing an increasing on the cumulative relative genetic contribution of ancestors

from 46.6% to 57.6%, indicating that the genetic base of Brazilian soybean is still narrow, despite the incorporation of new ancestors.

Table 2 - Sub-population structure with number of cultivar and selfing rates by group obtained for 169 cultivars of tropical soybean

Group	Number of cultivars	Selfing rates
1	16	0.962
2	25	0.964
3	17	0.965
4	16	0.966
5	14	0.966
6	28	0.967
7	19	0.967
8	18	0.968
9	16	0.970
Mean	-	0.966

Based on the alleles of 4,949 SNPs markers, and considering the nine subpopulations obtained with InStruct, the average molecular coancestry among the pairwise subpopulation comparisons was 0.234 in the tropical soybean collection as a whole. Approximately 60% of the pairwise coancestry estimates were lowest to 0.23 (Mean=0.196), 30% ranged from 0.24 to 0.3 (Mean=0.264), and 10% was higher than 0.31 (Mean=0.332) (Figure 3).

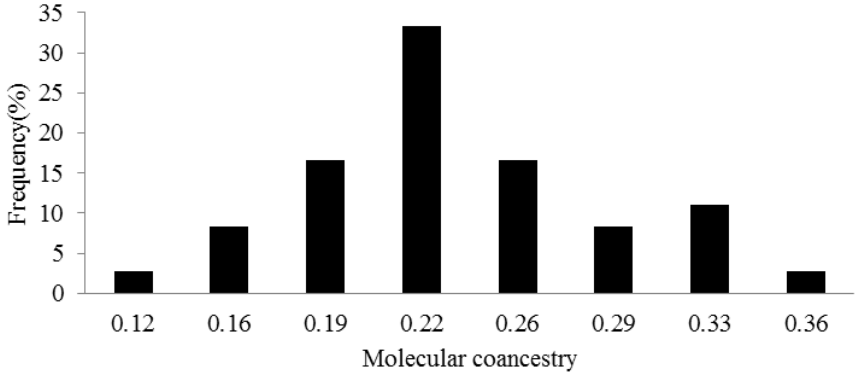


Figure 3 - Global pairwise molecular coancestry estimates of the 169 tropical soybean cultivars that represent nine subpopulations of Brazil.

Among the nine subpopulations, none had individuals exclusively from one company or maturity group (Tables 3 and 4). Company origin and MG may be the principal determinants of population structure within the soybean germplasm collection, however as the genetic base and origin of improved tropical lines are common it's difficult to explain it. Soybeans are classified into 13 unique MG Roman Numeral groups from very early to very late (000, 00, 0, I, II, III, IV, V, VI, VII, VIII, IX and X), based on temperature and photoperiod response to latitude. Our collection is represented by MG IV to IX and showed admixture population structure among the K nine subpopulations. Bandillo et al. (2015) evaluating a diverse soybean MG from the USDA Germplasm Resources Information Network (GRIN) database, reported that near of two-thirds of the accessions in the USDA soybean germplasm collection are admixed. Specifically, more than 90% of accessions from America and Europe are admixed. Probably it helps to confirm the admixed genetic structure nature of tropical soybean which has been developed from individuals that have a narrow genetic base of United States. In fact, previous studies demonstrate that the top five ancestors of Brazilian germplasm are the exact same top five ancestors for the soybean genetic base of the southern United States (Wysmierski and Vello, 2013).

Table 3 - Distribution of cultivars in each subgroup based on population structure and maturity groups of improved soybean tropical lines

MG	Clusters of instruct								
	1	2	3	4	5	6	7	8	9
IV	0	1	0	0	0	1	0	1	0
V	2	1	2	2	1	2	0	2	2
VI	8	11	6	5	5	10	9	12	4
VII	1	7	5	6	6	12	4	2	6
VIII	5	5	3	3	2	3	6	1	4
IX	0	0	1	0	0	0	0	0	0

The proportion of individuals for each company and MG within each of the nine subpopulations was not equals indicating different degrees of allelic diversity across populations, similar with the results reported by Bandillo et al. (2015) for the USDA soybean germplasm. As expected, individuals of each company of tropical soybean mostly were admixed in all subpopulations as a whole (Table 4).

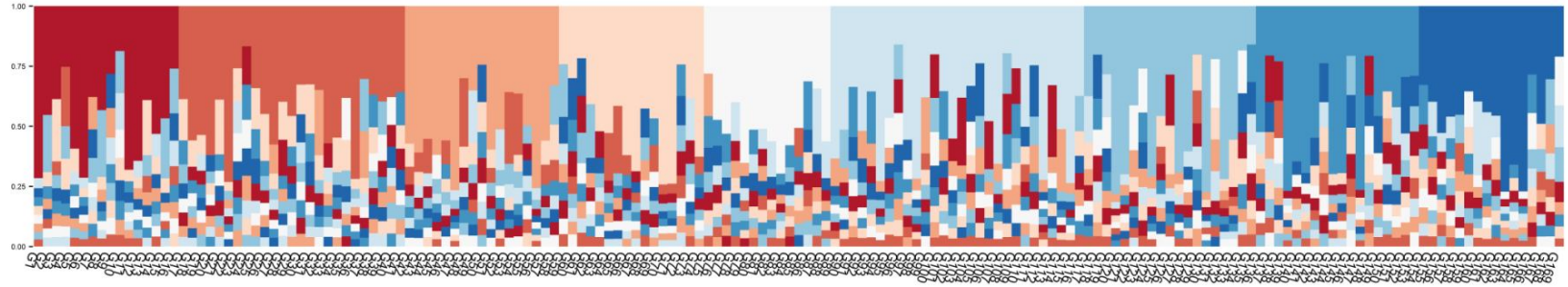


Figure 2 - Bar plot of the estimated population structure of 169 cultivars of soybean (k=9). The *y*-axis is the subgroup membership, and the *x*-axis is the genotype. The groups go from G1 to G9 from left to right.

Bandillo et al. (2015), indicate that the analysis of this result is complicated by the fact that ancestors of American soybean, the origin of most of the tropical soybean germplasm (Hiromoto and Vello, 1986), contributed at different pedigree levels, coupled with the fact that the American soybean germplasm resulted from a severe population bottleneck when soybeans were introduced to North America (Gizlice et al., 1994) and consequently to Brazil (Hiromoto and Vello, 1986). In consequence, company of origin and MG should be explaining a small genetic variation of tropical soybean.

Table 4 - Distribution of cultivars in each subgroup based on population structure and companies of improved soybean tropical lines

Company	Clusters of Instruct								
	1	2	3	4	5	6	7	8	9
DowAgroscience	0	0	0	1	1	2	2	0	0
GDM	2	1	3	1	1	0	0	0	1
Embrapa	5	7	7	1	2	5	5	5	6
Coodetec	6	9	2	5	6	10	5	6	6
Bayer	1	2	0	1	2	1	2	2	0
Igra	0	0	0	0	0	1	1	1	1
Monsanto	1	3	3	3	0	2	0	0	1
Syngenta	0	0	0	0	0	3	0	0	0
Nidera	0	1	0	2	0	1	1	1	0
Pionner	1	0	0	0	0	1	1	0	0
TMG	0	1	2	0	1	1	2	2	0
unknown	0	1	0	2	1	1	0	1	1

Hierarchical F statistics, calculated using ancestry estimates for K=9, showed that genetic differentiation explained by MG (~5%) was higher than that explained by Companies (~3%). Similar values of genetic differentiation for MG (MG 000 to X) using ancestry estimates for K=5 has been reported by Bandillo et al. (2015) for the USDA-GRIN soybean collection. Although the amount of total variation explained is small, these results suggest that population structure in the germplasm collection of Brazil is driven more by MG than companies of origin of soybean cultivars.

3.3. LD blocks analysis and LD-decay by chromosome

The SNPs with MAF > 0.1 distributed over the soybean genome (3,780) has permitted to identify 941 linkage disequilibrium blocks in the tropical soybean material, with 3,086 SNPs constituting the haplotype LD block (62% from total SNPs) (Table 1). In mean, the number of blocks by chromosome was 47.05, ranging among 32 (chromosome 1) to 74(chromosome 18) (Table1). The quantity of SNP in linkage disequilibrium in each block ranged between 2 to 9, with average of between 2 to 3 SNPs by block. Among the blocks in LD, 64% presented two or three markers, and less than 3% presented seven or more SNPs (Figure S1). The length of the blocks was very similar by chromosome, and most of these were represented among 51 to 500kb. Length blocks higher than 500kb was not found or was in a very low proportion. There was no relationship between the number of SNPs markers and the increase in linkage disequilibrium blocks, indicating that these blocks are randomly localized into the genome. The mean of the length of blocks was 252.4 kb, ranging among 1 (chromosome 4) to 499 kb (Chromosome 11). More than 70% of LD blocks showed a length lower than 200 kb (Figure S2). The sums of the lengths for LD blocks were 237,535 kb, and represents 20% of soybean genome, which have a length of 1.1 gb.

In this study LD decay was very high and variable among chromosomes (Table 5). At the moment, no information exists about the LD decay in improved tropical soybean lines adapted to Brazil with maturity groups among IV to IX. In addition, most of the studies conducted in soybean, has been used accessions from the U.S. Department of Agriculture (USDA) Soybean Germplasm Resources Information Network (GRIN) database (www.arsgrin.gov). In comparison with the GRIN soybean germplasm resource, with similar MG (Wen et al., 2015; Vuong et al., 2015) our improved tropical soybean showed a higher LD decay (Table 5). The difference of LD patterns may be attributed to low genome coverage of markers and fewer genotypes used in our study. Consequently, as suggested by Song et al. (2015) for soybean, most of the studies conducted for LD evaluation have been limited in terms of sample size and/or the number of loci analyzed, in fact, probably is necessary to evaluate the germplasm of tropical soybean with a greater number of markers.

Table 5 - Summary of LD decay rate (kb) comparison across 20 chromosomes within tropical soybean and improved soybean of U.S

Chr.	Improved tropical soybean	Improved U.S North-Central soybean*	USDA soybean germplasm**
1	8.4	226	250
2	5.4	276	300
3	4.8	135	150
4	1.2	113	200
5	9.7	270	300
6	12.5	206	175
7	4.5	235	500
8	1.3	242	250
9	3.5	190	250
10	10.2	158	200
11	4.9	176	200
12	12.9	175	250
13	3.9	311	200
14	5.2	317	300
15	9.4	305	400
16	27.9	101	125
17	4.3	171	225
18	4.2	375	500
19	5.2	430	600
20	2.1	259	150

* Maturity groups I, II and III (Wen et al., 2015). Landraces from multiple geographic origins including China, Japan, Korea, Kyrgyzstan and Russia, and materials developed for the U.S., North Central production area. LD decay at r^2=0.5.
**Maturity groups II, III, IV and V (Vuong et al., 2015). Collection represents 10 % of the total number of introduced soybean accessions in the USDA Soybean Germplasm Collection LD decay at r^2= 0.2.

We found that LD declined to r2 below 0.5 at ~2 Mbp (Figure 4) and it was variable among chromosomes, varying from 1.2kb (Chromosomes 4 and 8) to 27kb (Chromosome 16) (Figure 4). In improved cultivars that represent public and private breeding programs for the north central states of the United States (MG 0 and early

I), LD declined to r^2 below 0.1 at 7.0 Mbp, 5.9 Mbp and 8Mbp in the years 2005, 2006 and 2013, respectively (Mamidi et al., 2011, 2014). In Elite cultivars of a single breeding program of Canada r^2 dropped below 0.1 at ~2.8 Mb (Bastien et al., 2014). Hyten et al. (2007), reported a declined LD decay to an r^2 =0.1 at 574 kb in North American Elite Cultivars. In fact, highly variable pattern of LD have been reported in multiple soybean populations, and photoperiod sensitivity (maturity) has been proposed how a factor that may have contributed to increase LD in soybean, because their effect resulted in population subdivision in elite soybean cultivars (Hyten et al., 2007). Bastien et al. (2014), suggest that their results of less extensive LD is likely a reflection of the broader scope of the genotypes as it comprised genetically-modified, conventional, and food-type soybeans belonging to Maturity Groups 000 to II. In contrast, our tropical soybean collection showed high relationship among them, and this maybe explains our more extensive LD decay respect to others studies conducted with germplasm of soybean. It is not surprising to find high levels of LD in cultivars with high genetic relationship. In fact, the stringent cleistogamy and relatively long generation time of soybeans suggested that there would be high LD in the soybean genome (Lam et al., 2010).

It is known that LD increases with selfing and can extend very far in highly selfed organisms (Nordborg, 2000). Nordborg and Donnelly (1997) showed that the degree of selfing that a species exhibits is related to effective recombination rate. This is because recombination is less effective in selfing species where individuals are more likely to be homozygous at a given locus than in outcrossing species. In the current study, tropical soybean cultivars showed selfing rates equal to s=0.966 (data not shown). This relationship between recombination rate and selfing can extend to LD, because effective recombination is reduced severely in highly selfing species, as soybean, and consequently LD will be more extensive.

Cultivars contain specific sequence blocks in their chromosomes, which may be associated with artificially selected phenotypic variations from many generations of breeding (Kim et al., 2014; Song et al., 2015). The current study identified an extensive LD, with a set of 941 LD blocks, with most of the SNPs (3,086 or 62% from total SNPs number) constituting the haplotype LD blocks. Song et al. (2015) recently provided the first high-resolution haplotype maps based on the largest sample size and the largest number of loci reported in soybean thus far, and they identified that the extent of LD and the average haplotype block sizes were the greater in the North

American cultivar population, respect to wild and landraces populations. Our results were similar with the result reported for North American cultivars, and probably this corroborate that the extensive use of a small number of elite genotypes in Brazilian breeding program further reduces genetic variability. In fact, domestication and artificial selection have led to extensive LD and haplotype structure.

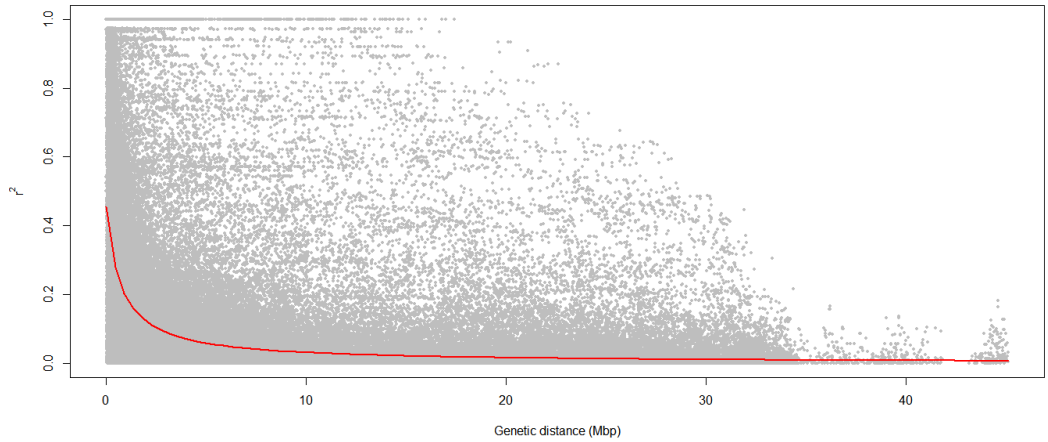


Figure 4 - LD decay among 169 cultivars of tropical soybean.

This study provides the first comprehensive sequencing data of tropical soybean genome and explored approximately 20% of soybean genome, considering that the sum of lengths for LD blocks were 237,535 kb. According to Schmutz et al. (2010), the soybean genome has about 1.1 to 1.15 gb, which means that this study was used one marker by 222 kb for evaluate the LD decay and LD haplotype blocks in tropical soybean. Our results showed small differences in length and number of LD blocks and demonstrate that the frequency of occurrence of LD blocks of lengths <500 kb is predominant in cultivated soybean of Brazil. Lam et al. (2010) reported that the frequency of occurrence of LD blocks of lengths <20 kb was higher in wild soybeans than in cultivated soybeans, and indicate that LD blocks of wild soybeans was about half that of cultivated soybeans. In fact, the genetic material used in this study maybe supported the relatively long LD blocks reported here.

Our results of high genetic relatedness and population structure in cultivars of tropical soybean, demonstrate that the nature of soybean fertilization, which results in high inbreeding and thus a reduction in recombination, may have promoted low genome diversity in the tropical soybean and high LD. According to Lam et al. (2010)

the presence of high LD in the soybean genome indicates that soybeans would serve as a good model for studying the genomes of crops with extreme LD. Additionally, the information provided by the present study about population structure, genetic relatedness and LD haplotype block location and distribution for cultivated soybean genome, can facilitate the identification of genes of interest. For breeding applications, our identification of the high LD nature in tropical soybean genome indicates that marker-assisted breeding and association mapping studies are better choices for soybean improvement, whereas mapbased cloning using genetic populations will be challenging.

4. REFERENCES

AKOND, M.; LIU, S.; SCHOENER, L.; ANDERSON, J.A.; KANTARTZI, S.K.; MEKSEM, K.; SONG, Q.; WANG, D.; WEN, Z.; LIGHTFOOT, D.A.; KASSEM, M.A. A SNP-based genetic linkage map of soybean using the SoySNP6K Illumina Infinium BeadChip genotyping array. **Journal of Plant Genome Science,** 1:80-89, 2013.

BANDILLO, N.; JARQUIN, D.; SONG, Q.; NELSON, R.; CREGAN, P.; SPECHT, J.; LORENZ, A. A Population Structure and Genome-wide Association Analysis on the USDA Soybean Germplasm Collection. **The Plant Genome,** 8:1-13, 2015.

BARRETT, J.C.; FRY, B.; MALLER, J.; DALY, M.J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics,** 21:263–265, 2005.

BASTIEN, M.; SONAHA, H.; BELZILEA, F. Genome Wide Association Mapping of *Sclerotinia sclerotiorum* resistance in soybean with a Genotyping-by-Sequencing Approach. **The Plant Genome,** 7:1-13, 2014.

BERNARDO, R.; ROMERO-SEVERSON, J.; ZIEGLE, J.; HAUSER, J.L.; HOOKSTRA, G.; DOERGE, R.W. Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. **Theoretical and Applied Genetics,** 100:552–556, 2000.

CARTER, T.E.; NELSON, R.; SNELLER, C.H.; CUI, Z. Genetic diversity in soybean. In: BOERMA, H.R.; SPECHT, J.E. (eds.). **Soybeans: Improvement, Production, and Uses**. Madison: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, 2004. p. 303–416.

EXCOFFIER, L.; LISCHER, H.E.L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Molecular Ecology Resources,** 10:564-567, 2010.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, 164:1567–1587, 2003.

FLINT-GARCIA, S.; THORNSBERRY, J.M.; BUKLER, E.S. Structure of linkage disequilibrium in plants. **Annual Review of Plant Biology,** 54:357–374, 2003.

FUJITA, R.; OHARA, M.; OKAZAKIA, K.; SHIMAMOTO, Y. The extent of natural crosspollination in wild soybean (*Glycine soja*). **Journal of Heredity,** 88:124–128, 1997.

GAI, J.Y.; XU, D.; GAO, Z.; SHIMAMOTO, Y.; ABE, J.; FUKUSHI, H.; KITAJIMA, S. Studies on the evolutionary relationship among eco-types of *G. max* and *G. soja* in China. **Acta Agronomica Sinica,** 26:513–520, 2000.

GAO, H.; WILLIAMSON, S.; BUSTAMANTE, C.D. A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. **Genetics,** 176:1635-1651, 2007.

GIZLICE, Z.; CARTER, T.E.; BURTON, J.W. Genetic base for North American public soybean cultivars released between 1947 and 1988. **Crop Science,** 34:1143-1151, 1994.

GORE, M.A.; CHIA, J.M.; ELSHIRE, R.J.; SUN, Q.; ERSOZ, E.S.; HURWITZ, B.L.; PEIFFER, J.A.; MCMULLEN, M.D.; GRILLS, G.S.; ROSS-IBARRA, J.; WARE, D.H.; BUCKLER, E.S. A first-generation haplotype map of maize. **Science,** 326:1115–1117, 2009.

GOUDET, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. **Molecular Ecology Notes,** 5:184-186, 2005.

HAUN, W.J.; HYTEN, D.L.; XU, W.W.; GERHARDT, D.J.; ALBERT, T.J.; RICHMOND, T.; JEDDELOH, J.A.; JIA, G.; SPRINGER, N.M.; VANCE, C.V.; STUPAR, R.M. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. **Plant Physiology,** 155:645-655, 2011.

HILL, W.G.; WEIR, B.S. Variances and covariances of squared linkage disequilibria in finite populations. **Theoretical Population Biology,** 33:54–78, 1988.

HIROMOTO, D.M.; VELLO, N.A. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazil Journal of Genetics,** 9:295-306, 1986.

HWANG, E.Y.; SONG, Q.; JIA, G.; SPECHT, J.E.; HYTEN, D.L.; COSTA, J.; CREGAN. P.B. A genome-wide association study of seed protein and oil content in soybean **BMC Genomics,** 2:1, 2014.

HYMOWITZ, T.; KAIZUMA, N. Soybean seed protein electrophoresis profiles from 15 Asian countries or regions: hypotheses on paths of dissemination of soybeans from China. **Economic Botany,** 35:10-23, 1981.

HYTEN, D.L.; CANNON, S.B.; SONG, Q.; WEEKS, N.; FICKUS, E.W.; SHOEMAKER, R.C.; SPECHT, J.E.; FARMER, A.D.; MAYA, G.D.; CREGAN, P.B. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. **BMC Genomics,** 15: 38, 2010.

HYTEN, D.L.; CHOI, I.Y.; SONG, Q.; SHOEMAKER, R.C.; NELSON, R.L.; COSTA, J.M.; SPECHT, J.E.; CREGAN, P.B. Highly variable patterns of linkage disequilibrium in multiple soybean populations. **Genetics,** 175:1937-1944, 2007.

HYTEN, D.L.; SONG, Q.; ZHU, Y.; CHOI, I.Y.; NELSON, R.L.; COSTA, J.M.; SPECHT, J.E.; SHOEMAKER, R.C.; CREGAN, P.B. Impacts of genetic bottlenecks on soybean genome diversity. **Proceedings of the National Academy of Sciences USA,** 103:16666–16671, 2006.

KIM, M.Y.; LEE, S.; VAN, K.; KIM, T.H.; JEONG, S.C.; CHOI, I.Y. KIM, D.S.; LEE, Y.S.; PARK, D.; MA, J.; KIM, W.Y.; KIM, B.C.; PARK, S.; LEE, K.A.; KIM, D.H.; KIM, K.H.; SHIN, J.H.; JANG, Y.E.; KIM, K.D.; LIU, W.X.; CHAISAN, T.; KANG, Y.J.; LEE, Y.H.; KIM, K.H.; MOON, J.K.; SCHMUTZ, J.; JACKSON, S.A.; BHAK, J.; LEE, S.H. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. **Proceedings of the National Academy of Sciences USA,** 107:22032-22037, 2010.

KIM, Y.H.; PARK, H.M.; HWANG, T.Y.; LEE, S.K.; CHOI, M.S.; JHO, S.; HWANG, S.; KIM, H.M.; LEE, D.; KIM, B.C.; HONG, C.P.; CHO, Y.C.; KIM, H.; JEONG, K.H.; SEO, M.J.; YUN, H.T.; KIM, S.L.; KWON, Y.U.; KIM, W.H.; CHUN, H.K.; LIM, S.J.; SHIN, Y.A.; CHOI, I.Y.; KIM, Y.S.; YOON, H.S.; LEE, S.H.; LEE, S. Variation block-based genomics method for crop plants. **BMC genomics,** 15:477, 2014.

LAM, H.M.; XU, X.; LIU, X.; CHEN, W.; YANG, G.; WONG, F.L.; LI, M.W.; HE, W.; QIN, N.; WANG, B.; LI, J.; JIAN, M.; WANG, J.; SHAO, G.; WANG, J.; SUNA, S.S.; ZHANG, G. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. **Nature Genetics,** 42:1053-1059, 2010.

LEE, S.; FREEWALT, K.R.; MCHALE, L.K.; SONG, Q.; JUN, T.H.; MICHEL, A.P.; DORRANCE, A.E.; MIAN, M.A.R. A high-resolution genetic linkage map of soybean based on 357 recombinant inbred lines genotyped with BARCSoySNP6K. **Molecular Breeding,** 35:58, 2015.

MAMIDI, S.; CHIKARA, S.; GOOS, R.J.; HYTEN, D.L.; ANNAM, D.; MOGHADDAM, S.M.; LEE, R.K.; CREGAN, P.B.; MCCLEAN, P.E. Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. **The Plant Genome,** 4:154-164, 2011.

MAMIDI, S.; LEE, R.K.; GOOS, J.R.; MCCLEAN, P.E. Genome-Wide Association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (*Glycine max*). **PLoS ONE,** 9:e107469, 2014.

NORDBORG, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. **Genetics,** 154:923–929, 2000.

NORDBORG, M.; DONNELLY, P. The coalescent process with selfing. **Genetics,** 146:1185-1195, 1997.

PRIOLLI, R.H.G.; CAMPOS, J.B.; STABELLINI, N.S.; PINHEIRO, J.B.; VELLO, N.A. Association mapping of oil content and fatty acid components in soybean. **Euphytica,** 203:83-96, 2014.

REMINGTON, D.L.; THORNSBERRY, J.M.; MATSUOKA, Y.; WILSON, L.M.; WHITT, S.R.; DOEBLEY, J.; KRESOVICH, S.; GOODMAN, M.M.; BUCKLER, E.S. Structure of linkage disequilibrium and phenotypic associations in the maize genome. **Proceedings of the National Academy of Sciences USA,** 98:11479-11484, 2001.

REYNOLDS, J.; WEIR, B.S.; COCKERHAM, C.C. Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. **Genetics,** 105:767-779, 1983.

SCHMUTZ, J.; CANNON, S.B.; SCHLUETER, J.; MA, J.; MITROS, T.; NELSON, W.; HYTEN, D.L.; SONG, Q.; THELEN, J.J.; CHENG, J.; XU, D.; HELLSTEN, U.; MAY, G.D.; YU, Y.; SAKURAI, T.; UMEZAWA, T.; BHATTACHARYYA, M.K.; SANDHU, D.; VALLIYODAN, B.; LINDQUIST, E.; PETO, M.; GRANT, D.; SHU, S.; GOODSTEIN, S.; BARRY, K.; FUTRELL-GRIGGS, M.; ABERNATHY, B.; DU, J.; TIAN, Z.; ZHU, L.; GILL, L.; JOSHI, T.; LIBAULT, M.; SETHURAMAN, A.; ZHANG, X.C.; SHINOZAKI, K.; NGUYEN, H.T.; WING, R.A.; CREGAN, P.; SPECHT, J.; GRIMWOOD, J.; ROKHSAR, D.; STACEY, G.; SHOEMAKER R.C.; JACKSON. S.A. Genome sequence of the palaeopolyploid soybean. **Nature,** 463:178-183, 2010.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. **Genes Genomes Genetics,** 1-17, 2015.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Development and evaluation of SoySNP50K, a High-density genotyping array for Soybean. **PLoS ONE,** 8:e54985, 2013.

SOTO-CERDA, B.; DIEDERICHSEN, A.; RAGUPATHYA, R.; CLOUTIER, S. Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. **BMC Plant Biology,** 13:78, 2013.

VIGOUROUX, Y.; MCMULLEN, M.; HITTINGER, C.T.; HOUCHINS, K.; SCHULZ, L.; KRESOVICH, S.; MATSUOKA, Y.; DOEBLEY, J. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. **Proceedings of the National Academy of Sciences USA,** 99:9650–9655, 2002.

VILLELA, O.T.; UNÊDA-TREVISOLI, S.H.; DA SILVA, F.M.; BÁRBARO, L.S.; DI MAURO, A.O. Genetic divergence of roundup ready (RR) soybean cultivars estimated by phenotypic characteristics and molecular markers. **African Journal of Biotechnology,** 13:2613-2625, 2014.

VUONG, T.D.; SONAH, H.; MEINHARDT, C.G.; DESHMUKH, R.; KADAM, S.; NELSON, R.L.; SHANNON, J.G.; NGUYEN, H.T. Genetic architecture of cyst

nematode resistance revealed by genome-wide association study in soybean. **BMC Genomics,** 16:593, 2015.

WEN, Z.; BOYSE, J.F.; SONG, Q.; CREGAN, P.B.; WANG, D. Genomic consequences of selection and genome-wide association mapping in soybean. **BMC Genomics,** 16:671, 2015.

WYSMIERSKI, P.T.; VELLO, N.A. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. **Genetics and Molecular Biology,** 36:547–555, 2013.

XU, D.H.; ABE, J.; GAI, J.Y.; SHIMAMOTO, Y. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. **Theoretical and Applied Genetics,** 105:645-653, 2002.

XU, D.H.; GAI, J.Y. Genetic diversity of wild and cultivated soybeans growing in China revealed by RAPD analysis. **Plant Breeding,** 122:503-506, 2003.

ZHOU, Z.; JIANG, Y.; WANG, Z.; GOU, Z.; LYU, J.; LI, W.; YU, Y.; SHU, L.; ZHAO, Y.; MA, Y.; FANG, C.; SHEN, Y.; LIU, T.; LI, C.; LI, Q.; WU, M.; WANG, M.; WU,Y.; DONG, Y.; WAN, W.; WANG, X.; DING, Z.; GAO, Y.; XIANG, H.; ZHU, B.; LEE, S.H.; WANG, W.; TIAN, Z. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. **Nature biotechnology,** 33:408-414, 2015.

CHAPTER 2 - A GENOME-WIDE ASSOCIATION STUDY FOR AGRONOMIC TRAITS IN SOYBEAN USING SNP MARKERS AND SNP-BASED HAPLOTYPE ANALYSIS

ABSTRACT

Mapping quantitative trait loci through the use of linkage disequilibrium (LD) in populations of unrelated individuals provides a valuable approach for dissecting the genetic basis of complex traits in soybean (*Glycine max*). The haplotype-based genome-wide association study (GWAS) has now been proposed as a complementary approach to intensify benefits from LD, which enable to assess the genetic determinants of agronomic traits. In this study a GWAS was undertaken to identify genomic regions that control 100-seed weight (SW), plant height (PH) and seed yield (SY) in a soybean association mapping panel using single nucleotide polymorphism (SNP) markers and haplotype information. The soybean cultivars (N = 169) were field-evaluated across four locations of southern Brazil. The genome-wide haplotype association analysis (941 haplotypes) identified eleven, seventeen and fifty-nine SNP-based haplotypes significantly associated with SY, SW and PH, respectively. Although most marker-trait associations were environment and trait specific, stable haplotype associations were identified for SY and SW across environments (i.e., haplotypes Gm12_Hap12). The haplotype block 42 on Chr19 (Gm19_Hap42) was confirmed to be associated with PH in two environments. These findings enable us to refine the breeding strategy for tropical soybean, which confirm that haplotype-based GWAS can provide new insights on the genetic determinants that are not captured by the single-marker approach.

Keywords: GWAS; haplotype-trait associations; linkage disequilibrium; yield-related traits.

1. INTRODUCTION

One of the most important crops for global production of vegetable protein and oil is Soybean (*Glycine max*). Due to quantitative inheritance of agronomic traits (seed protein, oil content and seed weight, for instance), several efforts have been made to understand the genetic basis of such complex traits (Hao et al., 2012; Zhang et al., 2014; Song et al., 2015; Zhang et al., 2015). Nowadays, with improved analytical methods for analyzing genome-wide association studies (GWAS), genomic selection (GS) and cost effective genotyping techniques there are promising forecasts in improving complex genetic traits in soybean (Zhang et al., 2014). In brief, GWAS use collections of diverse, unrelated lines that have been genotyped and phenotyped for certain traits of interest. Statistical associations between DNA polymorphism (or single nucleotide polymorphisms: SNP) are further investigated to identify genomic loci linked with a particular quantitative trait (Varshney et al., 2014). GWAS is useful to identify genes that code for important complex traits in crops such as those with self-pollinating mating systems (Lorenz et al., 2010). When compared to quantitative trait loci (QTL) studies that are achieved using pedigrees (e.g., bi-parental crosses), GWAS have the advantage of detecting smaller chromosomal regions affecting a trait and provides precise estimates of the size and direction of the effects of alleles in known loci (Abdel-Shafy et al., 2014). The natural genetic drift and random processes of mutations outcomes as linkage disequilibrium (LD) between markers and QTL where GWAS can benefit (Hamblin and Jannink, 2011). It has been seen that there is a high variable pattern of LD in soybean populations not only between populations but also in different regions of the genome (Hyten et al., 2007; Lam et al., 2010).

In order to enforce improvement in crops, SNP markers have turned out to be a potential tool in soybean breeding programs (Song et al., 2013; Zhang et al., 2014). SNP markers have also been employed in other important crops such as maize (Yan et al., 2011), rice (Yu et al., 2014) and wheat (Mora et al., 2015). SNP markers have enabled to improve the odds of success in a diversity of applications in soybean breeding programs, including positional cloning, association analysis, QTL mapping, and the determination of genetic relationships among individuals (Choi et al., 2007; Patil et al., 2016).

Looking at LD from an analytical point of view, it has been seen that it is best described using the haplotype-block approach (Hyten et al., 2010). The haplotype block is defined as a genomic region where a set of neighboring polymorphic loci (allelic variants) are in strong linkage disequilibrium in a population of interest (Greenspan and Geiger, 2004; Hamblin and Jannink, 2011). Hamblin and Jannink (2011) using coalescent simulations to compare single-SNP and haplotype markers, found that, across a range of plausible scenarios, the average power of 2- and 3-SNP haplotype markers to detect a QTL exceeds that of single-SNP markers. The specific haplotype blocks of soybean chromosomes can be associated with artificially selected phenotypic variations of many breeding generations (Kim et al., 2014) facilitating the identification of genes related with traits of interest (Lam et al., 2010).

It could be beneficial for GWAS to use haplotype information in making marker-phenotype associations (Lorenz et al., 2010) and could also compensate the bi-allelic limitation of SNP markers, and substantially improve the efficiency of QTL detection (Garner and Slatkin, 2003; Lu et al., 2010; Yan et al., 2011). In fact, according to Abdel-Shafy et al. (2014), GWAS using haplotype information in addition to using single-SNP could provide new insights on the genetic determinants that are not captured by the single-marker approach. Thus, the aim of this study was to identify genomic regions that control 100-seed weight (SW), plant height (PH) and seed yield (SY) in a soybean association mapping panel using individual SNP markers and haplotype information.

2. MATERIAL AND METHODS

2.1. Plant material and growing conditions

The association panel consisted of 169 genotypes that represent the core cultivars used by Brazilian farmers from 1990 to 2010, and some of these were key progenitors in soybean breeding programs of Brazil. The cultivars were field-evaluated in four sites of Brazil: Cascavel (24˚52'55"S 53˚32'30"W; 781asl), Palotina (24˚21'07"S 53˚45'25"W; 320asl), Primavera do Leste (15˚34'38"S 54˚20'42"W; 636asl) and Rio Verde (17˚45'49"S 51˚01'49"W; 330asl) (Table S1). Field trials were conducted using a randomized complete block design with two replicates. Fertilizer and field management practices recommended for optimum soybean production were used according to Embrapa (2011).

2.2. SNP genotyping

The cultivars were genotyped with 6,000 single nucleotide polymorphisms (SNP) using the Illumina BARCSoySNP6K BeadChip, which corresponds to a subset of SNPs from the SoySNP50K BeadChip (Song et al., 2013). Genotyping was conducted by Deoxi Biotechnology Ltda, in Aracatuba, Sao Paulo, Brazil. A total of 3,780 polymorphic and non-redundant SNP markers, with greater than 10% minor allele frequency (MAF) and missing data lower than 25% were used for subsequent analysis. Heterozygous markers were treated as missing data according to Hwang et al. (2014).

2.3. SNP-based haplotype blocks

941 haplotype blocks (characterized from the 3,780 SNPs) were used in this genome-wide association study. Haplotype blocks were constructed using the Solid Spine method implemented in the software Haploview (Barrett et al., 2005). This method considers that the first and last markers in a block are in strong LD with all intermediate markers, thereby providing more robust block boundaries. A cutoff of 1% was used, meaning that if addition of a SNP to a block resulted in a recombinant allele at a frequency exceeding 1%, then that SNP was not included in the block. The SNPs markers significantly associated with SY, PH and SW and located at the same

haplotype blocks were considered as a potential region of putative loci controlling the traits under study.

2.4. Population structure

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

$$\mathrm{DIC} = \overline{\mathrm{D}} + \mathrm{pD} \qquad (1)$$

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

2.5. Phenotypic data analysis

The following agronomic traits were measured and field-evaluated in the growing season 2012/2013: Seed yield (SY), 100-Seed Weight (SW) and Plant Height (PH). A mixed linear model was used for phenotypic data analysis using the MIXED procedure in SAS (SAS Institute, Inc., Cary, NC). The model that represents the combined data analysis was the following:

$$y_{ijk} = \mu + g_i + l_j + (gl)_{ij} + b_{k(j)} + e_{ijk} \qquad (2)$$

where μ is the total mean; g_i is the genetic effect of the i^{th} genotype; l_j is the effect of the j^{th} environment; $(gl)_{ij}$ is the interaction effect between the i^{th} genotype and the j^{th} environment (G × E); $b_{k(j)}$ is the random block effect within the j^{th} environment; and e_{ijk} is a random error following N(0, σ_e^2). Adjusted entry means (AEM) were calculated

for each of the 169 entries (i^{th} genotype: g_i) with the option LSMEANS of MIXED procedure, which were used as a dependent variable in the posterior association analysis (Mora et al., 2016). AEM denoted as M_i was:

$$M_i = \hat{\mu} + \hat{g}_i \qquad (3)$$

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

$$\mathrm{RLRT} = 2 \cdot \log\left[\frac{\mathrm{L}(\mathrm{M}_{\mathrm{HV}})}{\mathrm{L}(\mathrm{M}_{\mathrm{CV}})}\right] \qquad (4)$$

where M_{HV} and M_{CV} are the models with heterogeneous and common (homogenous) variances, respectively. The asymptotic distribution of the RLRT statistic is Chi-square with *p* degrees of freedom ($\mathrm{RLRT} \sim \chi^2_p$), where *p* is the difference in the number of parameters included in the M_{HV} and M_{CV} models (in this case *p*=3). Consequently, error variances were assumed to be heterogeneous among locations, which was computed using the REPEATED statement, option GROUP, of MIXED procedure.

Correlations among traits were determined following the method described by Holland et al. (2006), using the SAS macro (%macro correlation), which performs multivariate REML (Restricted Maximum Likelihood) estimation of variance and covariance components.

2.6. Association mapping analysis

AEM values were used to perform single-SNP analysis and then haplotype-based genome-wide association for the traits under consideration. In order to take into account the effects of population structure and genetic relatedness among the cultivars, the following unified mixed-model (Yu et al., 2006; Cappa et al., 2013) of association was employed (in matrix form):

$$y = S\alpha + Qv + Zu + \varepsilon \quad (5)$$

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

$$\mathrm{BIC} = -2 \cdot \log L + p \cdot \log(n) \quad (6)$$

where *L* is the restricted maximum likelihood for a determined model; *p* the number of parameters to be estimated in the model; and *n* the sample size. BIC values were computed using the TASSEL program following Yu et al. (2006). Haplotype-based association mapping was performed by using the Q + K model, following the unified mixed-model (Yu et al., 2006). A limit of detection (LOD) value higher than 3 was used as threshold P-value for both SNP- marker and haplotype-trait associations (Hwang et al., 2014). Then, only significant SNPs or haplotypes were used to estimate the phenotypic variance explained by the markers. The percent of variation explained by both SNP markers and SNP-based haplotypes was calculated by a regression analysis using TASSEL (Bradbury et al., 2007; Mamidi et al., 2014). The Chi-square test was performed to check phenotypic differences among haplotype blocks using the CONTRAST option of GENMOD procedure of SAS (SAS Institute, Inc., Cary, NC).

Additionally, the genomic regions (or SNPs in haplotypes blocks) identified in this study were compared to the genomic locations of QTLs previously reported for

the traits under study. Genes, QTLs and markers annotated in Glyma1.01 and NCBI RefSeq gene models in SoyBase (www.soybase.org) were used as reference.

3. RESULTS AND DISCUSSION

Analysis of variance indicated that the effects of genotype (G), environment (E) and their interaction (G × E) were statistically significant ($p < 0.01$) for all three traits under study (SY, SW and PH). This result is in agreement with the mixed model analysis, in which the 169 cultivars presented significant differences at $P < 0.01$ in all traits. The statistical results of fixed effects for the complex traits are summarized in Table 1. The mean seed yield (SY) varied significantly across locations. Soybean plants grown in Palotina had the lowest mean SY, while in Rio Verde plants had the highest SY. Plant height (PH) was significantly increased in Cascavel, while in Primavera do Leste PH was numerically decreased. However, plants in Primavera do Leste had the highest mean in 100-seed weight (SW).

Table 1 - Analysis of fixed effects for seed yield (SY, in $kg \cdot ha^{-1}$), plant height (PH, in cm) and 100-seed weight (SW, in g) measured in an association panel of soybean grown in four sites of southern Brazil. Data are presented as phenotypic means with standard deviations in parentheses

Trait	Environment				Mean squares		
	Cascavel	Palotina	Primavera	Rio Verde	E	G×E	G
SY	2322 (779)	1037 (381)	1890 (735)	2535 (839)	219490**	220491**	52737**
PH	104 (18)	89 (21)	49 (12)	57 (14)	32.6**	75.4**	158.3**
SW	12 (1.9)	11 (1.2)	13 (1.8)	12 (1.4)	0.78**	0.69**	1.36**

*Significant at the 0.01 probability level according to type III tests of fixed effects; G, genotype; E, environment; G×E, genotype-by-environment interaction.

Estimates of correlation coefficients among traits are shown in Table 2. SY was positively and significantly correlated with SW in three sites (estimates varied from 0.29 to 0.47; $P < 0.01$). The correlation estimate between SW and PH was not statistically different from zero, which was observed in all environments. On the other hand, there was no definite correlation between SY and PH; i.e., the correlation coefficient (calculated between these both traits) was negative in Cascavel, but positive in Primavera do Leste and Rio Verde.

Table 2 - Genotypic correlations among seed yield (SY), seed weigh (SW) and plant height (PH) in tropical soybean by environment

Environment	Trait	SY	SW
Cascavel	SW	0.47**	
	PH	-0.39**	-0.18ns
Palotina	SW	0.37**	
	PH	-0.02ns	-0.03ns
Primavera do Leste	SW	0.29**	
	PH	0.51**	-0.20ns
Rio Verde	SW	0.07ns	
	PH	0.54**	-0.49ns

** Significant at the 0.01 probability level; ns, not significant.

3.1. Population structure

In the present study, population structure of a soybean association panel consisting of 169 cultivars was investigated using a Bayesian clustering approach and a core set of SNP markers. According to the average log (likelihood) and the deviance information criterion (from the posterior Bayesian clustering analysis), the most probable number of subpopulations is nine. The probability of membership to each cluster indicates that 43% of all genotypes presented more than 50% of membership to their respective groups. However, most of them had an admixed condition. In fact, each subpopulation contained admixed cultivars that come from different soybean genetic breeding programs of Brazil (Table S1).

3.2. SNP-based association analysis

For model fit evaluation of mixed linear models with Q (structure) and K (kinship) matrices, the results based on Bayesian information criterion consistently showed a better fit for the (Q + K) model over the model that consider either Q or K alone (Table S2) for all data set (three traits and four environments). As shown in the quantile-quantile (QQ) plots (Figures S3-S8), the observed P-values from models that only include either population structure (Q model) or familial relatedness (K model), were significantly increased compared with the selected mixed model. Thus, the mixed linear model that includes Q and K (Q + K model) reduced the excess of low

P-values (Figures S3-S8). According to mixed-model analyses, six, seven and twenty-eight SNPs were significantly associated with SY, SW and PH, respectively (Tables S3, S4 and S5).

Six SNPs were significantly associated with SY on three chromosomes across two locations (Figures 1A and 1D), i.e., Cascavel (5) and Rio Verde (1). No significant SNPs were found in either Palotina nor Primavera do Leste (Figures 1B and 1C). The SNP ss715614920 associated with SY in Cascavel was identified on chromosome 13 at the intron region of the gene glyma13g25740, which encodes a putative germinal-center associated nuclear protein-like (Soybase, 2016) (Table S3).

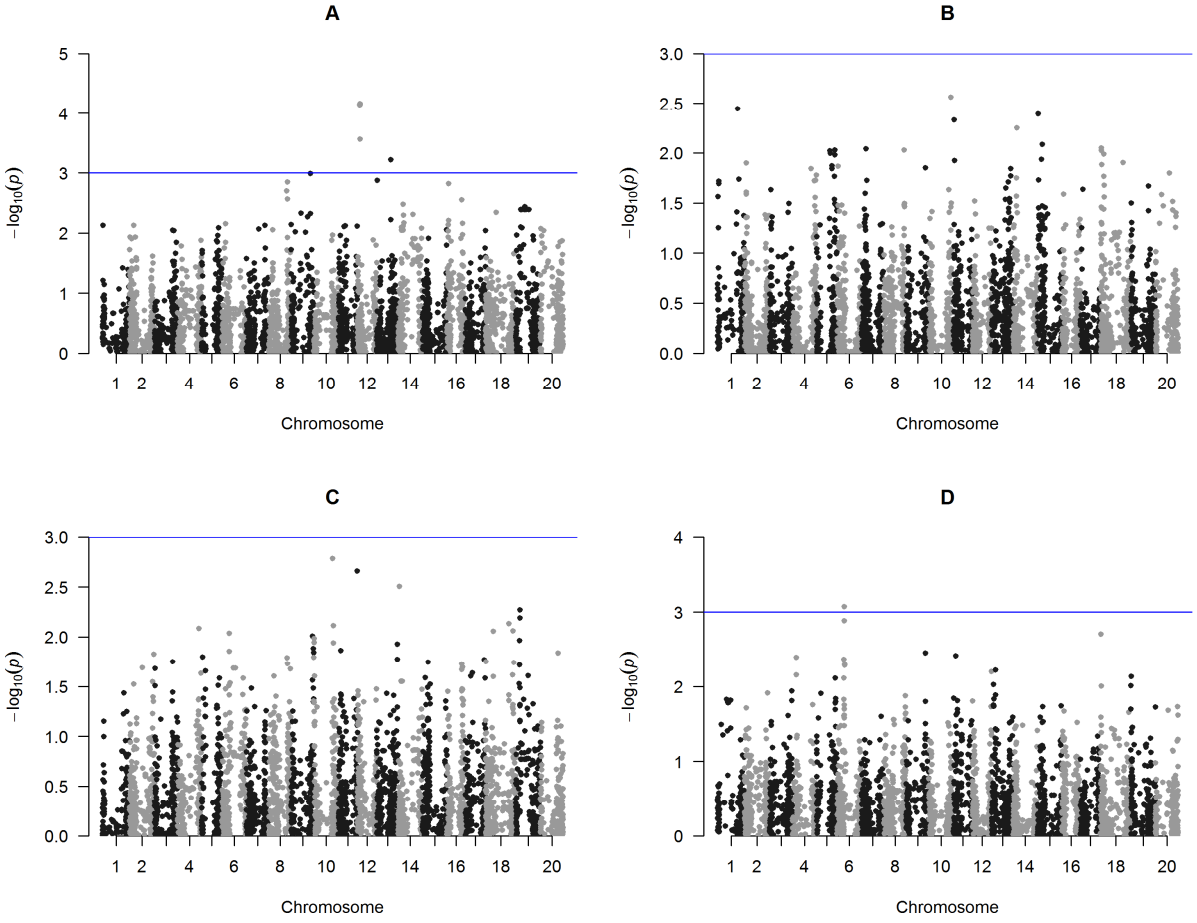


Figure 1 - Manhattan plots of GWAS for seed yield (SY) evaluated in a soybean association mapping panel across the following environments of southern Brazil, A) Cascavel, B) Palotina, C) Primavera do Leste and D) Rio Verde. Negative $\log_{10}$-transformed P-values of SNPs from a genome-wide scan for SY using a mixed linear model that includes both kinship and populations structure are plotted against positions on each of the 20 chromosomes. The significant SNPs associated with the trait ($P > 3.0 \times 10^{-3}$) are distinguished by the threshold line.

In Cascavel, the significant SNP ss715613203 (SY) was located in the same linkage disequilibrium block Gm12_Hap12 with the SNP ss715613192, ss715613207 and ss715613219. For this reason, this SNP is in linkage disequilibrium with the same genes and proteins associated with this LD block: Gm12_Hap12, i.e., uncharacterized gene LOC102667945 and the putative gene glyma12g075700 annotated as a double-stranded RNA-binding protein 2-like, which encodes a ribonuclease III protein (Figure 2, Tables 3 and 4). This LD block is also tightly linked to glyma12g075600, which encodes a senescence regulator in soybean. In addition, this LD block is close to markers satt568 and satt192 SSR, which have been involved in seed protein synthesis (Liang et al., 2010) and associated with QTLs of seed glycitein (Yang et al., 2001), respectively (Figure 2). The satt442 is a SSR marker located near to this haplotype region, which is associated to QTLs for seed protein, pod maturity and reproductive stage length in soybean. Importantly, this haplotype region has also been associated with SW in Palotina and Primavera do Leste in this study. The proportion of phenotypic variation explained by SNP-SY associations ranged from 9.14% (i.e., SNP ss715614920 located on Chr13 in Cascavel) to 15.83% (SNP ss715593323 on Chr6 in Rio Verde) (Table S3).

Seven SNPs were significantly associated with SW on chromosomes 5, 7, 11 and 12 across the locations under study (Figures 3A, 3B and 3C). In Cascavel, the two SNPs associated with SW (i.e., ss715592623 and ss715592632) are in a genomic region on Chr5 that encodes an elongation factor Ts mitochondrial-like (LOC100784416) and a ferredoxin-NAD(P) reductase activity protein (glyma05g09390), respectively (Soybase, 2016).

The SNPs of the Gm12_Hap12 were associated to SW in Palotina and Primavera do Leste (Tables 3 and 4). Other SNPs associated to SW in Primavera do Leste were: ss715598558 and ss715610817 located on chromosome 7 and 11, respectively. The SNP ss715598558 is located at the CDS region of the Glyma07g076800 gene, which encodes a transcription factor HEX, containing HOX and HALZ domains in soybean (Soybase, 2016). In Rio Verde, no SNP were found associated to SW (Table S3, Figure 3D).

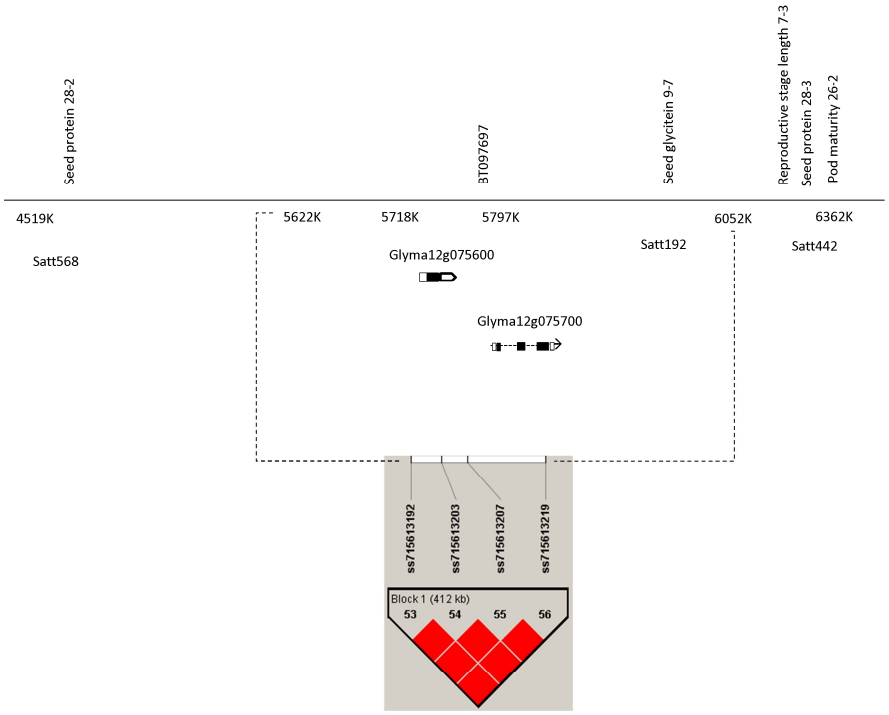


Figure 2 - Putative region (SNPs ss715613192 ss715613203, ss715613207 and ss715613219 on Gm12_Hap12) associated with seed weight (SW) and seed yield (SY) in soybean. Bottom panel depicts a haplotype region of 412 kb associated with SY and SW (Red color intensity indicates the intensity of r^2, i.e., higher color intensity means higher r^2).

One-hundred seed weight (SW) is one of the major yield components having direct effect on the final seed yield. For this trait, the proportion of phenotypic variance explained by a single genomic region found in this study was 9.92% in Cascavel (SNPs ss715592623 and ss715592632). In Palotina, the phenotypic variation ranged from 12.33% (ss715613104) to 13.31% (ss715613203). In Primavera do Leste, marker-SW associations explained from 8.92% (ss715613203) to 10.08% (ss715610817) of the phenotypic variation (Table S4).

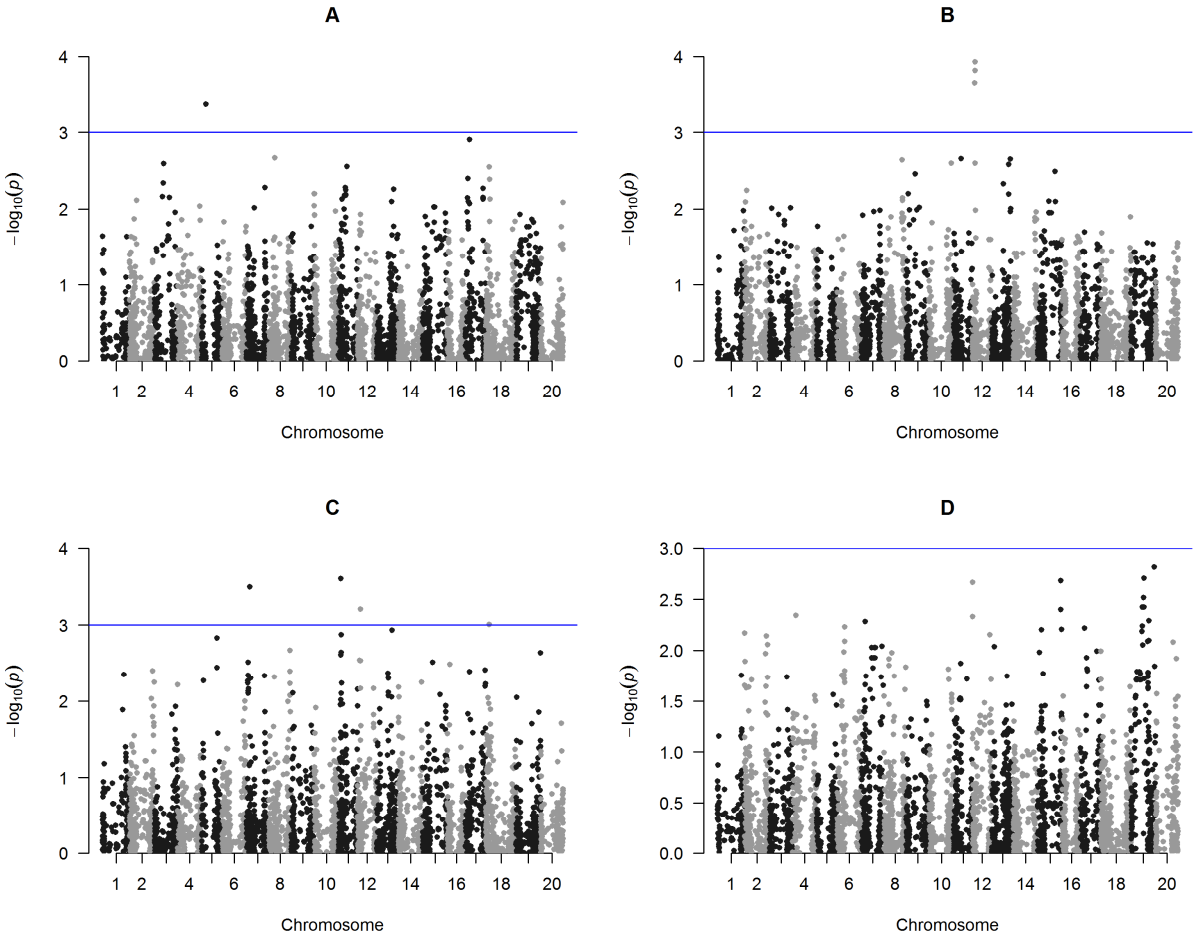


Figure 3 - Manhattan plots of GWAS for 100-seed weight (SW) evaluated in a soybean association mapping panel across the following environments of southern Brazil, A) Cascavel, B) Palotina, C) Primavera do Leste and D) Rio Verde. Negative $\log_{10}$-transformed P-values of SNPs from a genome-wide scan for SW using a mixed linear model that includes both kinship and populations structure are plotted against positions on each of the 20 chromosomes. The significant associations ($P > 3.0 \times 10^{-3}$) are distinguished by the threshold line.

Twenty-eight SNPs were significantly associated with PH across the four locations (Table S5), of which seventeen SNPs were found in Cascavel (Figure 4A), eleven in Palotina (Figure 4B), five in Primavera do Leste (Figure 4C) and three in Rio Verde (Figure 4D). The SNPs ss715601733, ss715609800, ss715581751 and ss715585767, which were associated to PH in Cascavel, showed no entry related with genes and/or molecular markers in the soybean database (Soybase, 2016). On the other hand, the SNPs ss715633774, ss715632400, ss715634905 and ss715622494 associated to PH in Cascavel, have been found in the same genomic regions that encode for development and cell death domain (glyma19g091100), heat

shock cognate 70 kDa protein 2-like, a heat stress transcription factor B-3-like and a cysteine synthase-like (glyma15g262500), respectively (Soybase, 2016).

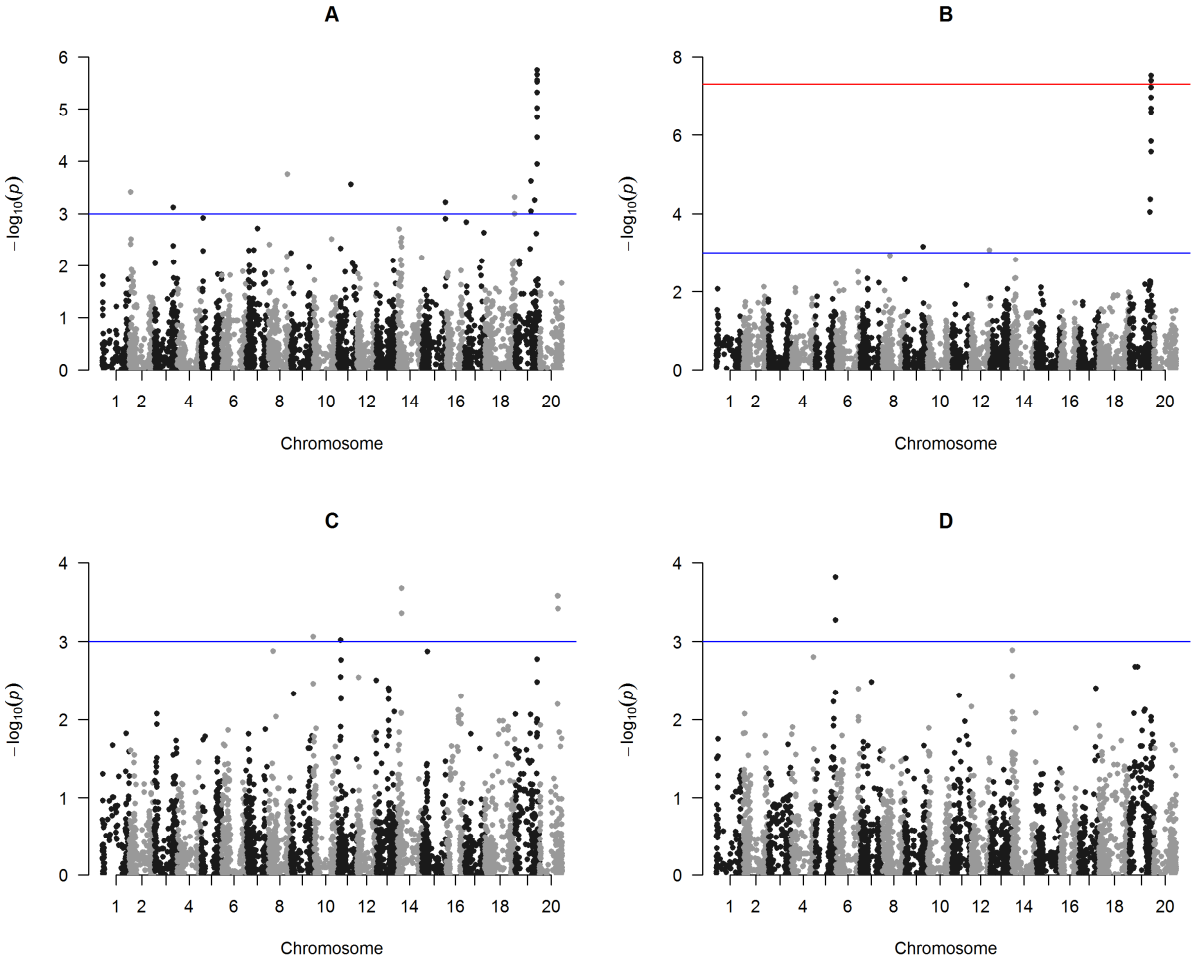


Figure 4 - Manhattan plot of GWAS for plant height (PH) evaluated in a soybean association mapping panel across the following environments of southern Brazil, A) Cascavel, B) Palotina, C) Primavera do Leste and D) Rio Verde. Negative $\log_{10}$-transformed P-values of SNPs from a genome-wide scan for PH using a mixed linear model that includes both kinship and populations structure are plotted against positions on each of the 20 chromosomes. The significant associations ($P > 3.0 \times 10^{-3}$) are distinguished by the threshold line.

In Palotina, the SNP markers ss715635224 and ss715603983, located on chromosomes 19 and 9, respectively, showed no entry with genes and/or molecular markers related to PH in soybean (Soybase, 2016). However, the SNP ss715635276, located on chromosome 19, is positioned close to a genomic region that encodes a heat shock cognate 70 kDa protein-like, as well as, other SNPs co-

associated with PH in Cascavel (Table S5). Similarly, the SNP ss715635468, identified on chromosome 19, showed strong significant association to PH in Cascavel and Palotina environments. In addition, it was related to glyma19g196000 gene, described as a probable UDP-N-acetylglucosamine-peptide N-acetylglucosaminyl transferase SPINDLY gene (S5 Table) (Soybase, 2016).

In Primavera do Leste, the SNP markers ss715619979, ss715637964 and ss715637991 were located on intergenic regions and showed no encoded genes related to plant height (Soybase, 2016). The same pattern was observed for the SNPs ss715592226 and ss715592231, which were associated to PH in Rio Verde. In contrast, the SNP markers ss715637988 and ss715619968 that were associated to PH in Primavera do Leste are on a genomic region that encodes an uncharacterized LOC100810047 (glyma20g28915) and a centromere-associated protein E-like (LOC100804944; glyma14g10050), respectively. Similarly, the genomic region on chromosome 5 (SNP ss715592240 associated to PH in Rio Verde) has been found to be involved to the synthesis of a probable protein S-acyltransferase 5-like (LOC100788304; glyma05g38360). In fact, the SNPs ss715592226 and ss715592231 were located in the same linkage disequilibrium block (Gm5_Hap40).

The haplotype block 42, associated to PH on Chr19 (Gm19_Hap42), is a region containing the *Determinate stem 1* gene (Dt1; Glyma19g37890) at 18.6 kb upstream of the peak SNP ss715635425 (Chr19_45000827; S5 Table and Table 6), which has been previously associated with PH and days to maturity in soybean [4] (Fig 5). In addition, other marker-yield associations have been previously identified at this region, seed yield 11-6, Plant height 13-8 and Plant height 4-2 (Lee et al., 1996; Specht et al., 2001) and associated with Dt1 (Cober et al., 2000) (Figure 5).

3.3. Haplotype blocks associated with complex traits

The genome-wide haplotype association analysis (941 haplotypes) identified eleven, seventeen and fifty-nine SNP-based haplotypes significantly associated with SY, SW and PH, respectively. As expected, both the size (kb) and the number of SNPs by LD block were highly variable (Tables 4, 5 and 6). Most of the blocks identified for each trait are in euchromatic regions according to the Glyma1.01 genome assembly.

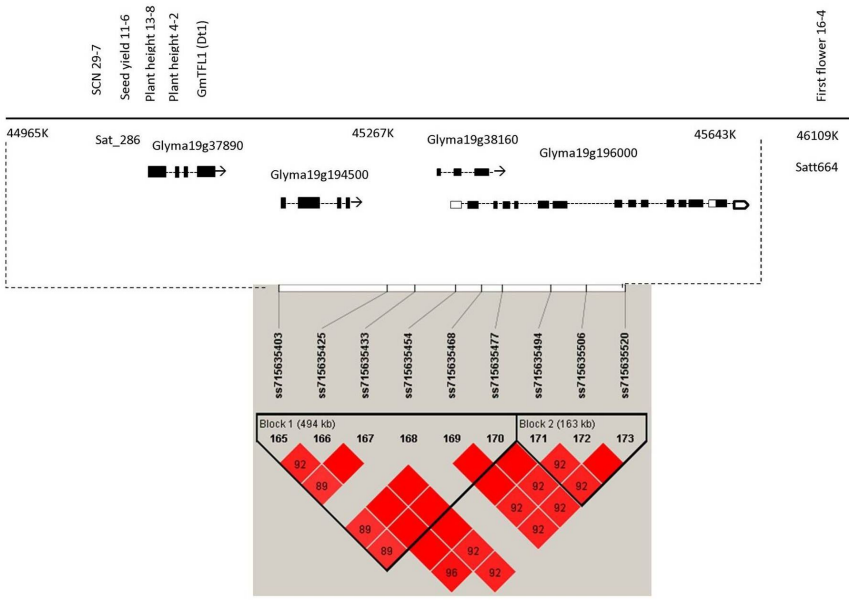


Figure 5 - Putative region (SNPs ss715635403, ss715635425, ss715635433, ss715635454 and ss715635468 located on Gm19_Hap42; and loci ss715635494, ss715635506 and ss715635520 located on Gm19_Hap43) associated to traits of interest. Gm19_Hap42 was associated with PH, SY and SCN in soybean. (Red color intensity indicates the intensity of r^2, i.e., higher color intensity means higher r^2).

For SY in Cascavel, the haplotypes TAAT (Gm12_Hap12a) and TAAC (Gm12_Hap12b) showed significant differences with the haplotype CGGT (Gm12_Hap12c). Gm12_Hap12a and Gm12_Hap12b had a mean value of 2567 kg ha^{-1} and 2381 kg ha^{-1}, respectively, while the haplotype Gm12_Hap12c yielded a mean of 1929 kg ha^{-1}, a yield 19% and 25% lower than the haplotypes Gm12_Hap12a and Gm12_Hap12b, respectively (Table 3). For SW, in Palotina, the same haplotypes (Gm12_Hap12a and Gm12_Hap12b) showed statistical differences with Gm12_Hap12c. In average, the haplotypes Gm12_Hap12a and Gm12_Hap12b had values of 11.4 g/100 seeds and 11.5 g/100 seeds, while the haplotype Gm12_Hap12c yielded 10.5 g/100 seeds (respectively 8% and 9% lower SW than Gm12_Hap12a and Gm12_Hap12b). In Primavera do Leste, the same haplotypes did not show statistical differences for SW. These haplotypes had the following frequencies in this association mapping panel: 30% for Gm12_Hap12a, 44% for Gm12_Hap12b and 26% for Gm12_Hap12c, and explained together a phenotypic variation of 12.1% for SY in Cascavel; 31.2% and 21.9% for SW in Palotina and Primavera do Leste, respectively (Tables 4 and 5).

Table 3 - Haplotypes associated with SY (mean in kg/ha) in 169 cultivars of tropical soybean

Environment	Chr	Position (bp) Start	Position (bp) End	Hap ID*	HapA*	HF[a]	R^2 (%)	SY [b]	Other nearby QTLs and genes[¶]
Cascavel	12	5610868	6023395	Gm12_Hap42a	TAAT	42	12.1	2566.5 a	Ribonuclease III satt568; satt442 and satt192**
				Gm12_Hap42b	TAAC	62		2380.3 a	
				Gm12_Hap42c	CGGT	36		1929.4 b	
	13	28918187	28957669	Gm13_Hap36a	CT	34	3.5	2436.5 a	Putative germinal-center associated nuclear protein-like
				Gm13_Hap36b	AT	74		2418.8 a	
				Gm13_Hap36c	AC	18		2136.4 ab	
				Gm13_Hap36d	CC	13		1725.9 ab	
Rio Verde	6	15115808	15242800	Gm6_Hap29a	CC	2	21.0	3508.0 a	-
				Gm6_Hap29b	TC	25		3305.6 a	-
				Gm6_Hap29c	CT	16		2761.6 a	-
				Gm6_Hap29d	TT	104		2446.4 b	-

* Hap ID = Haplotype identification; HapA=haplotype alleles.
[a] HF= Haplotype frequency: the number of cultivars with the respective haplotype.
[b] The average over the frequency of cultivars for each environment and the statistical difference.
** satt568 and satt442 from Liang et al. (2010), satt192 from Yang et al. (2011).
[¶] Genes nearby of the haplotype block.

Table 4 - Haplotypes associated with SW (mean in g/100 seed) in 169 cultivars of tropical soybean

Environment	Position (bp)			Hap ID	HapA*	HF[a]	R^2 (%)	SW [b]	Other nearby QTLs and genes[¶]
	Chr	Start	End						
Cascavel	5	9012813	9097414	Gm5_Hap10a	AA	19	13.8	12.5 a	glyma05g09390
				Gm5_Hap10b	GG	135		11.7 a	
Palotina	12	5610878	6023395	Gm12_Hap42b	TAAC	62	31.2	11.5 a	Ribonuclease III**
				Gm12_Hap42a	TAAT	42		11.4 a	satt568; satt442 and satt192
				Gm12_Hap42c	CGGT	36		10.5 b	
Primavera do Leste	11	5065170	5238788	Gm11_Hap13a	AA	76	13.2	11.8 a	-
				Gm11_Hap13b	GA	22		12.4 a	-
	7	6604493	7096376	Gm7_Hap13a	GGCGAGG	20	14.8	13.3 a	Glyma07g076800
				Gm7_Hap13b	GGCAAAT	2		12.7 a	
				Gm7_Hap13c	GGCAGAG	2		12.6 a	
				Gm7_Hap13d	AATAGAG	15		12.2 a	
				Gm7_Hap13e	AATAAAT	66		12.2 a	
				Gm7_Hap13f	GACAGAG	9		12.0 ab	
				Gm7_Hap13g	GGCAAGG	19		11.8 abc	
	12	5610878	6023395	Gm12_Hap42b	TAAC	62	21.8	12.8 a	-
				Gm12_Hap42a	TAAT	42		12.3 a	-
				Gm12_Hap42c	CGGT	36		11.9 a	-

* Hap ID = Haplotype identification; HapA=haplotype alleles. [a] HF= Haplotype frequency: the number cultivars with the respective haplotype. [b] The average over the frequency of cultivars for each environment and the statistical difference. ** satt568 and satt442 [¶] Genes nearby of the haplotype block.

A discriminant haplotype was identified in a low frequency for PH in this association mapping panel, i.e. the haplotype Gm19_Hap42b in which the plants had a mean of 83.8 and 85.0 cm of height in Cascavel and Palotina, respectively. In both environments, this haplotype showed statistical difference with the haplotype responsible for produce tallest plants (Gm19_Hap42a). Together, these haplotypes explained a phenotypic variation of 91.4% and 96% in Cascavel and Palotina, respectively (Table 5). Another interesting genomic region was located on Chr9 (Gm9_Hap24), in which the haplotypes did not show statistical differences for PH, and the plants had a mean of 94.8cm (Gm9_Hap24a), 89.9cm (Gm9_Hap24b) and 87.5cm (Gm9_Hap24c) of height in Palotina. The haplotypes together explained 12% of the phenotypic variation for PH (Table 5).

3.4. GWAS and model selection

This study was undertaken to identify genomic regions associated with key complex traits in soybean. An advantage of using a genetically broad panel is the opportunity to explore alleles that could potentially be used in a marker-assisted selection context to improve agronomic traits in soybean. In fact, this GWAS approach employed the optimal mixed model identified valuable SNPs that were significantly associated with SY, SW and PH. In addition, to refine the association with SNPs markers, a haplotype-based analysis was performed to discover if these genomic regions were localized at the same haplotype blocks, and Williams 82 physical map. The soybean whole-genome sequence of SoyBase (2016) provided key insights about sequence-based genetic markers, previously reported as significant for these traits in soybean.

Genetic relatedness (or kinship) and population structure are known as the major confounding factors that may lead to spurious associations in GWAS (Yu et al., 2006). In consequence, we tested all MLMs with the combination of Q and K matrices. The Q + K model consistently fit the best according to BIC and -2*log*L, compared with either Q or K models. A lower inflation of P-values was consistently observed when Q + K models were employed in data analyses. This analytical model has been recognized as an effective model to perform genome-wide association for complex traits in many plant species (Hwang et al., 2014; Song et al., 2015), which has allowed accurate analysis of association studies in soybean (Zhang et al., 2015).

Table 5 - Haplotypes associated with PH (Mean in cm) in 169 cultivars of tropical soybean

Environment	Chr	Position (bp) Start	End	Hap ID	HapA*	HF [a]	R^2 (%)	PH [b]	Other nearby QTLs and genes[¶]
Cascavel	19	44761515	45255796	Gm19_Hap42a	AATxAA	34	91.4	111.62 a	Sd yld 11-6 *
				Gm19_Hap42b	GCCGGG	110		101.18 b	Pl ht 4-2
									Pl ht 13-8
				Gm19_Hap42c	ACCGGG	2		83.75 b	Glyma19g196000
									Glyma19g37890 Dt1 gene
	19	45361938	45525374	Gm19_Hap43a	GTA	2	44.1	121.25 a	-
				Gm19_Hap43b	ATA	34		112.28 a	
				Gm19_Hap43c	GCG	111		100.98 ab	
				Gm19_Hap43d	ACG	2		90.00 ab	
	19	32194361	32318695	Gm19_Hap20a	CG	57	17.3	107.68 a	LOC100789162
				Gm19_Hap20b	TA	87		105.53 a	
	18	61175038	61450878	Gm18_Hap71a	ATGG	7	22.2	115.36 a	LOC100787543
				Gm18_Hap71b	ATAT	76		109.13 ab	
				Gm18_Hap71c	ATAG	15		108.67 abc	
				Gm18_Hap71d	GCGG	31		99.44 abc	
				Gm18_Hap71e	GTGG	9		94.70 bc	
	19	39686084	40143590	Gm19_Hap34a	TGAT	13	9.1	108.65 a	LOC100786140
				Gm19_Hap34b	TGGC	3		107.50 a	
				Gm19_Hap34c	CGGC	23		107.28 a	
				Gm19_Hap34d	TTAT	25		101.40 a	
				Gm19_Hap34e	TTGC	70		100.38 a	

Table 5, cont.

	15	48653554	48727813	Gm15_Hap45a	CC	81	109.33 a	LOC100804065
				Gm15_Hap45b	AC	6	105.00 a	
				Gm15_Hap45c	AT	64	100.08 ab	
				Gm15_Hap45d	CT	2	90.00 ab	
	3	38761991	38976026	Gm3_Hap32a	TAAT	51	108.87 a	-
				Gm3_Hap32b	GGCT	29	105.26 a	
				Gm3_Hap32c	GGCC	49	104.92 a	
				Gm3_Hap32d	GGAT	4	100.63 a	
	19	44761515	45255796	Gm19_Hap42a	AATxAA	34	107.03 a	-*
				Gm19_Hap42b	ACCGGG	2	85.00 ab	
				Gm19_Hap42c	GCCGGG	110	78.33 b	
Palotina	19	45361938	45525374	Gm19_Hap43b	ATA	34	106.88 a	-
				Gm19_Hap43a	GTA	2	105.00 ab	
				Gm19_Hapd	ACG	2	80.00 ab	
				Gm19_Hapc	GCG	111	78.75 abc	
	19	42812863	43117852	Gm19_Hap38a	TA	29	106.34 a	LOC100777767
				Gm19_Hap38b	TC	7	83.00 b	
				Gm19_Hap38c	CC	113	82.30 bc	

Table 5, cont.

Primavera	9	38013391	38454149	Gm9_Hap24a	AA	59	12.0	94.87 a	-
				Gm9_Hap24b	GG	53		89.89 a	
				Gm9_Hap24c	GA	24		87.50 a	
	14	8027761	8527621	Gm14_Hap21a	CGGGTA	4	47.6	63.75 a	LOC100804944
				Gm14_Hap21b	CGGGGA	37		55.39 a	
				Gm14_Hap21c	CGTATA	8		52.25 a	
				Gm14_Hap21d	TTTAGA	19		51.15 ab	
				Gm14_Hap21e	TTTATA	47		48.00 ab	
				Gm14_Hap21f	CGTAGA	2		46.25 abc	
				Gm14_Hap21g	TTTAGG	14		41.46 bc	
	20	37857633	38195568	Gm20_Hap24a	GGxTG	16	27.6	66.56 a	LOC100810047
				Gm20_Hap24b	AATTG	2		57.50 a	
				Gm20_Hap24c	AATTA	78		47.25 ab	
				Gm20_Hap24d	AATCG	2		44.75 b	
	20	37211061	37410040	Gm20_Hap23a	GC	14	19.3	67.44 a	-
				Gm20_Hap23b	AT	140		48.04 b	
				Gm20_Hap23c	GT	2		44.75 c	
Rio Verde	5	41481303	41866018	Gm5_Hap40a	TCCCG	3	55.3	70.00 a	LOC100788304
				Gm5_Hap40b	CCCCG	45		69.48 ab	
				Gm5_Hap40c	TTTTG	47		56.25 b	
				Gm5_Hap40d	TTTTA	33		52.86 b	
				Gm5_Hap40e	CTTTG	2		.	

* HapID = Haplotype identification; HapA=haplotype alleles.
[a] HF= Haplotype frequency; the number of cultivars with the respective haplotype.
[b] The average over the frequency of cultivars for each environment and its statistical difference.
** Also associated in Palotina; Pl ht 13-8 and Pl ht 4-2 from Lee et al. (1996); Sd yld 11-6 from Specht et al. (2001); Dt1 gene from Cober et al. (2000). [¶] Genes nearby of the haplotype block.

3.5. Correlation among traits

SY had a positive and significant correlation with SW, which is in agreement with previous reports in soybean (Hao et al., 2012; Recker et al., 2014). The undefined correlation between SY and PH (significant positive and negative values) observed in this study, has the same behavior as seen in previous studies (Kim et al., 2012; Fox et al., 2015). According to Kim et al. (2012) there is no consistent pattern in the relationship between seed yield and other important agronomic traits in soybean, but it has been shown that a generally higher yield is associated with later maturity and taller plant height (Coincibido et al., 2003; Kabelka et al., 2004; Fox et al., 2015).

3.6. Haplotypes and genomic regions associated with complex traits

Many studies have demonstrated the power of GWAS to detect significant QTL in soybean populations. In this study, we highlight the importance of having haplotype maps of tropical soybean cultivars for marker-assisted selection (MAS). Moreover, according to Lorenz et al. (2010) GWAS may benefit from utilizing haplotype information for making marker-phenotype associations and, in addition to the individual-SNP approach, offers further advantages for the molecular genetic dissection of loci underlying complex traits in soybean. Song et al. (2015) stated that with the advent of the haplotype block map, one could efficiently select SNPs and haplotypes blocks for optimized association analysis. In this study, notably, the haplotype Gm12_Hap12 showed a significant positive association with both SY and SW. Furthermore, the positive significant correlation between both traits may be a result of either genes in LD or genetic pleiotropy. Given the high association of few likely putative genomic regions, we could hypothesize that pleiotropic gene effects underlie the observed significant positive genotypic correlation between these traits. However, the reverse is also true, i.e., several SY and SW QTLs were identified independently (and localized on different genomic regions), evidencing the complexity of these traits. The possibility of coexistence of multiple genes should not be excluded due to the quantitative nature of the genetic background. Moreover, the sizes of the haplotype Gm12_Hap12 is greater than 412 Kb. Additionally, SNP markers co-associated with two or more traits at the same haplotype coincided with significant phenotypic and genotypic correlation among the studied traits, as reported

before (Hao et al., 2012; Kabelka et al., 2004). In soybean, MAS of a co-associated genetic locus could simultaneously improve multi-associated target traits, but additional studies are always necessary because the distinction between LD and pleiotropy will allow breeders to develop effective breeding methodologies to select and obtain favorable trait combinations (Recker et al., 2014).

Yield QTLs identified on chromosome 12 are of particular interest because they showed consistent effects across locations (Palotina, Primavera do Leste and Cascavel). Zhang et al. (2016) recently reported a close SNP (ss715613104) as effectively associated to SW in soybean. Furthermore, the following SSR markers: satt568, satt442 and satt192, which are linked to seed protein (Liang et al., 2010) and seed glycitein (Yang et al., 2011), respectively, have been co-localized near to the haplotype block identified on chromosome 12. One of the primary advantages of GWAS is the high mapping resolution. This feature enables GWAS to further narrow down the chromosomal region of putative QTLs and predict causal genes (Zhang et al., 2016). Biologically important genes were identified on this haplotype block region (Gm12_Hap12). The gene Glyma12g075700, which encodes a ribonuclease III protein, represents an uncharacterized protein associated to BT097697 code in SoyBase (2016). Glyma12g075600 is another gene located near to Gm12_Hap12, which encodes a protein for senescence regulator in soybean (i.e., annotated as a U-box domain-containing protein 13-like; phytozome.jgi.doe.gov/). Importantly, its homolog in *Arabidopsis thaliana* regulates the expression of proteins associated with leaf senescence (Fischer-Kilbienski et al., 2010).

The SNP at 45 Mb on Chr19 associated with PH has been previously reported by Lee et al. (1996) and Specht et al. (2001), which has QTLs associated with Seed Yield 11-6, Plant height 4–2 and Plant Height 13–8. Zhang et al. (2015) also reported this SNP, which was strongly associated to PH and days to maturity. In fact, this result indicated that some causal gene(s) might exist in this genomic region. These associated markers may be useful for aggregation of causal genes of interest to improve soybean yield. Furthermore, in this region some markers have been reported near to the Dt1 gene (Glyma19g37890) (Zhang et al., 2015). Dt1 is homologous to *Arabidopsis* terminal flower 1, and plays a predominant role in determining stem growth habit in soybean (Liu et al., 2010). Stem growth habit is an important discriminant trait for soybean cultivars classifying it in two major categories, determinate and indeterminate. Given the high relationship between plant growth

habit, plant height and seed yield in soybean, our result is highly consistent with the result of Zhang et al. (2015), who determined that the locus harboring Dt1 was strongly associated with PH.

Near to Dt1 gene, in the same haplotype Gm19_Hap42, was located the SPINDLY gene (SPY) (Glyma19g196000), which is considered to be a negative regulator of gibberellin (GA) signaling in *Arabidopsis thaliana*. Swain et al. (2002) proposed that the SPY gene acts independently of GA responses in controlling cotyledon number, leaf growth, hypocotyl growth and plant height. In our GWAS, this result makes sense because SPY was co-localized with genes of plant height and near to QTL controlling first flowering in soybean.

3.7. QTL x environment interaction

The significant G × E interaction explains the relatively low stability (or consistency) of the identified loci. Moreover, this result is important, because clearly justifies the inclusion of different environments (locations) in the GWAS. In fact, to obtain the real QTL with genetic stability and high phenotypic variation explained, different environments of the same material, QTL mapping and QTL geographic interactions should be used and explored (Sun et al., 2012). Due to the presence of a significant G × E interaction, QTL analysis was separately carried out in each location. In this study, most of the SNP-trait associations were location specific. When genotype or haplotype refers to QTL, this phenomenon is called QTL-by-environment interaction, denoted by Q × E (Zhao et al., 2012). The existence of Q × E reported here confirmed the complexity of the quantitative traits under study.

Only three SNPs (ss715613203, ss715613104 and ss715613207) and one haplotype (Gm12_Hap12) were detected to be stable for SY and SW with high correlation between these two traits in the four environments under consideration, which was due to that agronomic traits are the result of the combined actions of multiple genes and environmental factors; with gene expression varying across environments (Mansur et al., 1993). The inheritance of quantitative traits undeniably involves multiple genes with small effect that are sensitive to environmental changes (Xing and Zhang, 2010). The stable associations found in this study should be useful for the breeding purpose to find broad adaptability to different environments. In Brazil, the development of elite cultivars has long challenged breeders due to the

effects of large differences in latitude, climate, altitude, diversity of soil type, farming and planting practices, plant growth habit, presence or absence of long-juvenile traits, different stress conditions and diseases, resulting in large G × E interactions (Alliprandini et al., 2009). Thus, the marker-assisted selection using markers identified in a specific environment could be beneficial for breeders that attempt to identify the best landraces that are specifically adapted to local growing conditions.

In conclusion, with the aid of the haplotype block map constructed by Song et al. (2015) and our haplotype block results, we efficiently tested SNPs and SNP-based haplotypes for optimized association analyses. Importantly, various haplotypes were significantly associated with SY (11), SW (17) and PH (59), of which some were located in/or near regions where QTLs for yield and yield-related traits have been previously mapped by either linkage or GWAS analysis. Moreover, new haplotypes-trait associations have been identified in this study (as the case of Gm12_Hap12: Gm12_Hap12a and Gm12_Hap12b), which could be used as putative regions for further research efforts focusing on the genetic basis of soybean yield and yield components. These haplotypes showed the best performance in comparison with the Gm12_Hap12c haplotype, and depended upon both geographic location and traits.

Some haplotypes contain SNP markers that were not detected in the single-marker analysis (i.e., SY: Gm13_Hap36; SW: Gm7_Hap13 and Gm12_Hap12; PH: Gm14_Hap21). This is attributed to the nature of the haplotype-based method, which can better detect functional haplotypes such as *cis*-interactions among multiple DNA variants in a haplotype block region (Liu et al., 2008), and identify co-associated haplotype regions with two or more traits, indicating pleiotropy of single causal gene or tight linkage of multiple causal genes (Hao et al., 2012), which is an advantage of the haplotype analysis compared to the single SNP analysis. Another advantage of the haplotype-based method is that the small size of the haplotype regions (as identified in this study) would facilitate the search for causal genetic variations that affect gene functions, as stated by Abdel-Shafy et al. (2014).

The use of SNPs associated with quantitative trait loci under the allelic combination approach, for example, can be further used for the efficient marker assisted selection of complex traits (Mamidi et al., 2014). Moreover, the practical use of the haplotype identified in this study may contribute to increase the efficiency of the current breeding programs carried out in tropical regions worldwide. The results confirm that the haplotype-based GWAS provides new insights on the genetic

determinants that are not captured by the single-SNP approach. However, as any molecular markers, we emphasized that the identified haplotypes should be validated before large-scale use (Schuster, 2011).

Although SNP chips with higher density and next-generation sequencing may provide new data (Sonah et al., 2015), the results of this study suggest that BARCSoySNP6K BeadChip is a valuable source of information to discover genomic regions that control quantitative traits. Finally, this research identified useful haplotypes that have not been previously reported, which would help to assess and validate causal genetic variation of complex quantitative traits and eventually may be used for breeding purposes in soybean.

4. REFERENCES

ABDEL-SHAFY, H.; BORTFELDT, R.H.; TETENS, J.; BROCKMANN, G.A. Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. **Genetics Selection Evolution,** 46:35, 2014.

ALLIPRANDINI, L.F.; ABATTI, C.; BERTAGNOLLI, P.F.; CAVASSIM, J.E.; GABE, H.L.; KUREK, A.; MATSUMOTO, M.N.; DE OLIVEIRA, M.A.R.; PITOL, C.; PRADO, L.C.; STECKLING, C. Understanding soybean maturity groups in Brazil: environment, cultivar classification, and stability. **Crop Science,** 49:801-808, 2009.

BARRETT, J.C.; FRY, B.; MALLER, J.; DALY, M.J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics,** 21:263-265, 2005.

BRADBURY, P.J.; ZHANG, Z.; KROON, D.E.; CASSTEVENS, T.M.; RAMDOSS, Y.; BUCKLER, E.S. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics,** 23:2633-2635, 2007.

CAPPA, E.P.; EL-KASSABY, Y.A.; GARCIA, M.N.; ACUÑA, C.; BORRALHO, N.M.; GRATTAPAGLIA, D.; MARCUCCI POLTRI, S.N. Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. **PLoS ONE,** 8:e81267, 2013.

CHOI, I.Y.; HYTEN, D.L.; MATUKUMALLI, L.K.; SONG, Q.; CHAKY, J.M.; QUIGLEY, C.V.; CHASE, K.; LARK, K.G.; REITER, R.S.; YOON, M.S.; HWANG, E.Y.; YI, S.I.; YOUNG, N.D.; SHOEMAKER, R.C.; VAN TASSEL, C.P.; SPECHT, J.E.; CREGAN, P.B. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. **Genetics,** 176:685–696, 2007.

COBER, E.R.; MADILL, J.; VOLDENG, H.D. Early tall determinate soybean genotype E1E1e3e3e4e4dt1dt1 sets high bottom pods. **Canadian Journal of Plant Science,** 80: 527-531, 2000.

CONCIBIDO, V.C.; LA VALLEE, B.; MCLAIRD, P.; PINEDA, N.; MEYER, J.; HUMMEL, L.; YANG, J.; WU, K.; DELANNAY, X. Introgression of a quantitative trait

locus for yield from *Glycine soja* into commercial soybean cultivars. **Theoretical and Applied Genetics,** 106:575–582, 2003.

EMBRAPA. **Tecnologias de produção de soja – região central do Brasil 2012 e 2013**. - Londrina: Embrapa Soja, 2011. 261p. (Sistemas de Produção / Embrapa Soja, n.15)

ENDELMAN, J.B.; JANNINK, J.C. Shrinkage estimation of realized relationship matrix. **Genes Genomes Genetics,** 2:1405-1413, 2012.

FISCHER-KILBIENSKI, I.; MIAO, Y.; ROITSCH, T.; ZSCHIESCHE, W.; HUMBECK, K.; KRUPINSKA, K. Nuclear targeted AtS40 modulates senescence associated gene expression in *Arabidopsis thaliana* during natural development and in darkness. **Plant Molecular Biology,** 73:379-390, 2010.

FOX, C.M.; CARY, T.R.; NELSON, R.L.; DIERS, B.W. Confirmation of a Seed Yield QTL in Soybean. **Crop Science,** 55:992-998, 2015.

GAO, H.; WILLIAMSON, S.; BUSTAMANTE, C.D. A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. **Genetics,** 176:1635-1651, 2007.

GARNER, C.; SLATKIN, M. On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. **Genetic Epidemiology,** 24:57-67, 2003.

GREENSPAN, G.; GEIGER, D. Model-based inference of haplotype block variation. **Journal of Computational Biology,** 11:493-504, 2004.

HAMBLIN, M.T.; JANNINK, J.L. Factors affecting the power of haplotype markers in association studies. **The Plant Genome,** 4:145-153, 2011.

HAO, D.; CHENG, H.; YIN, Z.; CUI, S.; ZHANG, D.; WANG, H.; YU, D. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. **Theoretical and Applied Genetics,** 124:447-458, 2012.

HOLLAND, J.B. Estimating genotypic correlations and their standard errors using multivariate restricted maximum likelihood estimation with SAS Proc MIXED. **Crop Science,** 46:642-654, 2006.

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

HYTEN, D.L.; CHOI, I.Y.; SONG, Q.; SHOEMAKER, R.C.; NELSON, R.L.; COSTA, J.M.; SPECHT, J.M.; CREGAN, P.B. Highly variable patterns of linkage disequilibrium in multiple soybean populations. **Genetics,** 175:1937-1944, 2007.

KABELKA, E.A.; DIERS, B.W.; FEHR, W.R.; LEROY, A.R.; BAIANU, I.C.; YOU, T.; NEECE, D.J.; NELSON, R.L. Putative alleles for increased yield from soybean plant introductions. **Crop Science,** 44:784-791, 2004.

KIM, K.S.; DIERS, B.W.; HYTEN, D.L.; MIAN, M.A.R.; SHANNON, J.G.; NELSON, R.L. Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. **Theoretical and Applied Genetics,** 125:1353–1369, 2012.

KIM, Y.H.; PARK, H.M.; HWANG, T.Y.; LEE, S.K.; CHOI, M.S.; JHO, S.; HWANG, S.; KIM, H.M.; LEE, D.; KIM, B.C.; HONG, C.P.; CHO, Y.C.; KIM, H.; JEONG, K.H.; SEO, M.J.; YUN, H.T.; KIM, S.L.; KWON, Y.U.; KIM, W.H.; CHUN, H.K.; LIM, S.J.; SHIN, Y.A.; CHOI, I.Y.; KIM, Y.S.; YOON, H.S.; LEE, S.H.; LEE, S. Variation block-based genomics method for crop plants. **BMC genomics,** 15:477, 2014.

LAM, H.M.; XU, X.; LIU, X.; CHEN, W.; YANG, G.; WONG, F.L.; LI, M.W.; HE, W.; QIN, N.; WANG, B.; LI, J.; JIAN, M.; WANG, J.; SHAO, G.; WANG, J.; SUN, S.S.; ZHANG, G. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. **Nature Genetics,** 42:1053-1059, 2010.

LEE, S.H.; BAILEY, M.A.; MIAN, M.A.R.; JR CARTER, T.E.; ASHLEY, D.A.; HUSSEY, R.S.; PARROTT, W.A.; BOERMA, H.R. Molecular markers associated with soybean plant height, lodging, and maturity across locations. **Crop Science,** 36:728–735, 1996.

LIANG, H.; YU, Y.; WANG, S.; LIAN, Y.; WANG, T.; WEI, Y.; GONG, P.; LIU, X.; FANG, X.; ZHANG, M. QTL mapping of isoflavone, oil and protein contents in soybean (*Glycine max* L. Merr.). **Agricultural Science in China,** 9:1108-1116, 2010.

LIU, B.; WATANABE, S.; UCHIYAMA, T.; KONG, F.; KANAZAWA, A.; XIA, Z.; NAGAMATSU, A.; ARAI, M.; YAMADA, T.; KITAMURA, K.; MASUTA, C.; HARADA, K.; ABE, J. The soybean stem growth habit gene Dt1 is an ortholog of *Arabidopsis* TERMINAL FLOWER1. **Plant Physiology,** 153:198-210, 2010.

LIU, N.; ZHANG, K.; ZHAO, H. Haplotype-association analysis. **Advances in Genetics,** 60:335-405, 2008.

LORENZ, A.J.; HAMBLIN, M.T.; JANNINK, J.L. Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome-Wide Association Analysis in Barley. **PLoS ONE,** 5:e14079, 2010.

LU, Y.; ZHANG, S.; SHAH, T.; XIE, C.; HAO, Z.; LI, X.; FARKHARI, M.; RIBAUT, J.; CAO, M.; RONG, T. Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. **Proceedings of the National Academy of Sciences USA,** 107:19585–19590, 2010.

MAMIDI, S.; LEE, R.K.; GOOS, J.R.; MCCLEAN, P.E. Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (*Glycine max*). **PLoS ONE,** 9:e107469, 2014.

MANSUR, L.; LARK, K.; KROSS, H.; OLIVEIRA, A. Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). **Theoretical and Applied Genetics,** 86:907–913, 1993.

MORA, F.; CASTILLO, D.; LADO, B.; MATUS, I.; POLAND, J.; BELZILE, F.; VON ZITZEWITZ, J.; DEL POZO, A. Genome-wide association mapping of agronomic traits and carbon isotope discrimination in a worldwide germplasm collection of spring wheat using SNP markers. **Molecular Breeding,** 35:69, 2015.

MORA, F.; QUITRAL, Y.A.; MATUS, I.; RUSSELL, J.; WAUGH, R.; DEL POZO, A. SNP-based QTL mapping of fifteen complex traits in barley under rain fed and well-

watered conditions by a mixed modeling approach. **Frontiers in Plant Science,** 7:909, 2016.

PATIL, G.; DO, T.; VUONG, T.D.; VALLIYODAN, B.; LEE, J.D.; CHAUDHARY, J.; SHANNON, J.G.; NGUYEN, H.T. Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. **Scientific Reports,** 6:19199, 2016.

RECKER, J.R.; BURTON, J.W.; CARDINAL, A.; MIRANDA, L. Genetic and Phenotypic Correlations of Quantitative Traits in Two Long-Term, Randomly Mated Soybean Populations. **Crop Science,** 54:939–943, 2014.

SCHMUTZ, J.; CANNON, S.B.; SCHLUETER, J.; MA, J.; MITROS, T.; NELSON, W.; HYTEN, D.L.; SONG, Q.; THELEN, J.J.; CHENG, J.; XU, D.; HELLSTEN, U.; MAY, G.D.; YU, Y.; SAKURAI, T.; UMEZAWA, T.; BHATTACHARYYA, M.K.; SANDHU, D.; VALLIYODAN, B.; LINDQUIST, E.; PETO, M.; GRANT, D.; SHU, S.; GOODSTEIN, S.; BARRY, K.; FUTRELL-GRIGGS, M.; ABERNATHY, B.; DU, J.; TIAN, Z.; ZHU, L.; GILL, L.; JOSHI, T.; LIBAULT, M.; SETHURAMAN, A.; ZHANG, X.C.; SHINOZAKI, K.; NGUYEN, H.T.; WING, R.A.; CREGAN, P.; SPECHT, J.; GRIMWOOD, J.; ROKHSAR, D.; STACEY, G.; SHOEMAKER R.C.; JACKSON. S.A. Genome sequence of the palaeopolyploid soybean. **Nature,** 463:178-183, 2010.

SCHUSTER, I. Marker-assisted selection for quantitative traits. **Crop Breeding and Applied Biotechnology,** S1:50-55, 2011.

SCHWARZ, G. Estimating the dimension of a model. **Annals Statistics,** 6:461-464, 1978.

SONAH, H.; O'DONOUGHUE, L.; COBER, E.; RAJCAN, I.; BELZILE, F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. **Plant Biotechnology Journal,** 13:211-221, 2015.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Fingerprinting soybean germplasm and its utility in genomic research. **Genes Genomes Genetics,** 5:1999-2006, 2015.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE,** 8:e54985, 2013.

SOYBASE. (2016) **USDA-ARS Soybean genetics and genomics database**. USDA, Washington, DC. Disponível em: www.soybase.org/search/qtllist.php. Acesso em: 15, abril, 2016.

SPECHT, J.E.; CHASE, K.; MACRANDER, M.; GRAEF, G.L.; CHUNG, J.; MARKWELL, J.P.; GERMANN, M.; ORF, J.H.; LARK, K.G. Soybean response to water: a QTL analysis of drought tolerance. **Crop Science,** 4:493-509, 2001.

SPIEGELHALTER, D.J.; BEST, N.G.; CARLIN, B.P.; VAN DER LINDE, A. Bayesian measures of model complexity and fit, (with discussion). **Journal of the Royal Statistical Society Series B,** 64:583-639, 2002.

STICH, B.; MOHRING, J.; PIEPHO, H.P.; HECKENBERGER, M.; BUCKLER, E.S.; MELCHINGER, A.E. Comparison of mixed-model approaches for association mapping. **Genetics,** 178:1745-1754, 2008.

SUN, Y.N.; PAN, J.B.; SHI, X.L.; DU, X.Y.; WU, Q.; QI, Z.M.; JIANG, H.W.; XIN, D.W.; LIU, C.Y.; HU, G.H.; CHEN, Q.S. Multi-environment mapping and meta-analysis of 100-seed weight in soybean. **Molecular Biology Reports,** 39:9435-9443, 2012.

SWAIN, S.M.; TSENG, T.S.; THORNTON, T.M.; GOPALRAJ, M.; OLSZEWSKI, N.E. SPINDLY is a nuclear-localized repressor of gibberellin signal transduction expressed throughout the plant. **Plant Physiology,** 129:605-615, 2002.

VARSHNEY, R.K.; TERAUCHI, R.; MCCOUCH, S.R. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. **PLoS Biology,** 12:e1001883, 2014.

XING, Y.; ZHANG, Q. Genetic and molecular bases of rice yield. **Annual Review of Plant Biology,** 61:421–442, 2010.

YAN, J.; WARBURTON, M.; CROUCH, J. Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. **Crop Science,** 51:433-449, 2011.

YANG, K.; MOON, J.; JEONG, N.; CHUN, H.; KANG, S.; BACK, K.; JEONG, S. Novel major quantitative trait loci regulating the content of isoflavone in soybean seeds. **Genetics and Genomics,** 33:685-692, 2011.

YU, H.; XIE, W.; LI, J.; ZHOU, F.; ZHANG, Q. A whole-genome SNP array (RICE6K) for genomic breeding in rice. **Plant Biotechnology Journal,** 12:28-37, 2014.

YU, J.; PRESSOIR, G.; BRIGGS, W.; VROH, B.I.; YAMASAKI, M.; DOEBLEY, J.; MCMULLEN, M.; GAUT, B.; NIELSEN, D.; HOLLAND, J.; KRESOVICH, S.; BUCKLER, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics,** 38:203–208, 2006.

ZHANG, J.; SONG, Q.; CREGAN, P.B.; JIANG, G.L. Genome wide association study, genomic prediction and marker assisted selection for seed weight in soybean (*Glycine max*). **Theoretical and Applied Genetics,** 129:117-130, 2016.

ZHANG, J.; SONG, Q.; CREGAN, P.B.; NELSON, R.L.; WANG, X.; WU, J.; JIANG, G.L. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. **BMC genomics,** 16:217, 2015.

ZHAO, F.; XU, S. Genotype by environment interaction of quantitative traits: a case study in barley. **Genes Genomes Genetics,** 2:779-788, 2012.

CHAPTER 3 - A HAPLOTYPE-BASED GENOME-WIDE ASSOCIATION STUDY FOR FLOWERING, MATURITY DATES AND YIELD-RELATED TRAITS ACROSS MULTIPLE ENVIRONMENTS IN SOYBEAN

ABSTRACT

Knowledge of the genetic architecture for flowering and maturity is needed to develop effective breeding strategies in tropical soybean. The aim of this study was to identify haplotypes across multiple environments that contribute to flowering time and maturity, with the purpose of selecting desired alleles that have a minimal impact on yield-related traits in tropical soybean. A genome-wide association study (GWAS) was undertaken to identify genomic regions that control days to flowering (DTF), maturity (DTM) and yield-related traits (100-seed weight: SW, plant height: PH, seed yield: SY) using a soybean association mapping panel genotyped for single nucleotide polymorphism (SNP) markers. The soybean cultivars (N=141) were field-evaluated across eight tropical environments of Brazil. Our results revealed significant associations of 33, 29, 57, 72 and 40 SNP-based haplotypes with SY, SW, PH, DTM and DTF, respectively, in two or more environments. Haplotype-based GWAS identified three haplotypes (Gm12_Hap12; Gm19_Hap42 and Gm20_Hap32) significantly co-associated with DTF, DTM and yield-related traits in single and multiple environments. These results indicate that these genomic regions may contain genes that have pleiotropic effects controlling yield-related traits and flowering time and maturity and are tightly linked with multiple other genes with high rates of linkage disequilibrium.

Keywords: Dt1 gene; Allelic haplotypes; Pleiotropy.

1. INTRODUCTION

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

Yield-quantitative trait loci (QTL) are often detected within the context of specific soybean breeding populations and environments, since some conditions in any given environment, geographic region or year can change the grain yield (Orf et al., 1999; Guzman et al., 2007). According to Palomeque et al. (2010), studies have identified QTLs associated with traits of interest that appear to be independent of the environment but dependent on the genetic background in which they found. The difficulty of identifying yield-QTL effective for MAS across a wide range of genetic and/or environmental contexts might be addressable by using preliminary yield trials to model target haplotypes within each context and then immediately selecting inbred lines that target genotypes in real time (Sebastian et al., 2010). Sebastian et al. (2010) demonstrated that using MAS with haplotypes to improve grain yield is possible if focused within a specific genetic and environmental context. In addition, the context-specific approach has already been adopted as a major component of MAS strategies known commercially as Accelerated Yield Technology (AYT) at Pioneer Hi-Bred International.

Genome-wide association studies (GWAS) using individual Single Nucleotide Polymorphism (SNPs) and haplotype information have been used to improve agronomical traits in soybean (Hao et al., 2012; Zhang et al., 2015; Contreras-Soto et al., 2017). A haplotype block is a genomic region in which two or more polymorphic loci (i.e., SNP) in close proximity tend to be inherited together with high probability (Abdel-Shafy et al., 2014). These blocks are believed to be caused by recombination hotspots with extremely rare recombination within stretches of DNA, where the

enclosed SNPs consequently segregate together from one generation to the next, acting as combined multi-site alleles (Greenspan and Geiger, 2004). The combination of SNP alleles in a haplotype block on one chromosome covers the observed variation and can have higher linkage disequilibrium (LD) with the allele of a QTL than individual SNP alleles that are used to construct the haplotype (Abdel-Shafy et al., 2014). Furthermore, haplotype association is likely to be more powerful in the presence of LD (Garner and Slatkin, 2003). Lorenz et al. (2010) used simulated phenotype data to show that the use of SNP-based haplotypes can increase power over the use of single-SNP markers in GWAS. Using haplotypes for QTL mapping could compensate for the bi-allelic limitation of SNPs, and substantially improve the efficiency of QTL mapping (Yang et al., 2011). According to Song et al. (2015), highly selfing species, such as soybean, are in many ways uniquely suitable for haplotype block mapping. Therefore, the aim of this study was to identify haplotypes across multiple environments that contribute to time to flowering, maturity and plant height, to improve the selection of desired alleles with a minimal impact on yield in tropical soybean.

2. MATERIAL AND METHODS

2.1. Plant material and field evaluation

The association panel of this study consisted of 141 cultivars of tropical soybean, which were field evaluated in five locations that represent eight environments of Brazil: Cascavel (24°52'54.9"S 53°32'30.4"W) in the growing seasons 2012/2013, 2013/2014 and 2014/2015 (Env1, Env3 and Env5, respectively); Palotina (24°21'06.5"S 53°45'24.9"W) in the growing season 2014/2015 (Env6); Primavera do Leste (15°34'37.6"S 54°20'41.8"W) in the growing season 2012/2013 (Env2), Rio Verde (17°45'49.0"S 51°01'49.3"W) in the growing season 2013/2014 and 2014/2015 (Env4 and Env7), and Sorriso (12°32'43.6"S 55°42'41.8"W) in the growing season 2014/2015 (Env8). These locations were chosen on the basis of their diversity of latitude and altitude. Field trials were arranged in a complete block design with two replications. Fertilizer and field management practices recommended for optimum soybean production were used according to Embrapa (2011).

2.2. Phenotypic data analysis

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

$$y_{ijk} = \mu + g_i + l_j + (gl)_{ij} + b_{k(j)} + e_{ijk} \qquad (1)$$

where μ is the total mean, g_i is the genetic effect of the i^{th} genotype, l_j is the effect of the j^{th} environment, $(gl)_{ij}$ is the interaction effect between the i^{th} genotype and the j^{th} environment (G × E), $b_{k(j)}$ is the random block effect within the j^{th} environment, and e_{ijk} is a random error following $N(0, \sigma_e^2)$. Adjusted entry means (AEM) were calculated

for each of the 141 entries (i^{th} genotype: g_i) with the LSMEANS option of MIXED procedure, and these were used as a dependent variable in the posterior association analysis. AEM (denoted as M_i) was:

$$M_i = \hat{\mu} + \hat{g}_i \qquad (2)$$

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

$$\mathrm{RLRT} = 2 \cdot \log\left[\frac{\mathrm{L(M_{HV})}}{\mathrm{L(M_{CV})}}\right] \qquad (3)$$

where MHV and MCV are the models with heterogeneous and common (homogenous) variances, respectively. The asymptotic distribution of the RLRT statistic is Chi-square with p degrees of freedom ($\mathrm{RLRT} \sim \chi^2_p$), where p is the difference in the number of parameters included in the MHV and MCV models (in this case P=7). Consequently, error variances were assumed to be heterogeneous among locations, and these were computed using the COVTEST homogeneity option, with RANDOM _residual_ statement and GROUP option in the GLIMMIX procedure (Mora et al., 2016). Analysis of Deviance (ANODEV) was conducted to evaluate the significance of the effects of the five traits across environments by using the MIXED procedure in SAS (Nelder and Wedderburn, 1972). The PROC CORR procedure was used to analyze Pearson correlations among variables by environment since G × E interactions were significant. Broad-sense heritability for the five traits was estimated according to Hao et al. (2012).

2.3. Association panel, SNP genotyping and population structure

Cultivars were genotyped for 6,000 single nucleotide polymorphisms (SNPs) using the Illumina BARCSoySNP6K BeadChip, corresponding to a subset of SNPs from the SoySNP50K BeadChip (Song et al., 2013). Genotyping was conducted by Deoxi Biotechnology Ltda. ® in Aracatuba, Sao Paulo, Brazil. A total of 3,780 SNP markers, including polymorphic and non-redundant SNPs, SNP markers with greater than 10% minor allele frequency (MAF) and missing data values lower than 25% were used for subsequent analysis, with heterozygous markers treated as missing data. Haplotype blocks were constructed using the Solid Spine method implemented in the software Haploview (Barrett et al., 2005), and have been previously reported by Contreras-Soto et al. (2017).

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

$$\mathrm{DIC} = \bar{\mathrm{D}} + \mathrm{pD} \qquad (4)$$

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

2.4. Association mapping analysis

AEM calculated for each cultivar were used to perform haplotype-based genome-wide association for SY, PH, SW, DTF and DTM. To consider the effects of population structure and genetic relatedness among the cultivars, the following

unified mixed-model (Yu et al., 2006; Cappa et al., 2013) of association was employed (in matrix form):

$$y = S\alpha + Qv + Zu + \varepsilon \quad (5)$$

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

$$\mathrm{BIC} = -2 \cdot \log L + p \cdot \log(n) \quad (6)$$

where *L* is the restricted maximum likelihood for a determined model, *p* the number of parameters to be estimated in the model, and *n* the sample size. BIC values were computed using the TASSEL program following Yu et al. (2006). Haplotype-based association mapping was performed by using the Q + K model, following the unified mixed-model (Yu et al., 2006). A limit of detection (LOD) value higher than 3 was used as threshold P-value for haplotype-trait associations according to Hwang et al. (2014). Then, only the significant haplotypes were used to estimate the phenotypic variance explained by haplotypes. The percent of variation explained by the haplotype-based method was calculated using a simple regression performed in TASSEL (Bradbury et al., 2007). The Chi-square test was performed to check phenotypic differences among haplotype blocks using the CONTRAST option of GENMOD procedure in SAS (SAS Institute, Inc., Cary, NC).

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

3. RESULTS AND DISCUSSION

3.1. Phenotypic analysis, heritability and correlation between traits

Analysis of deviance indicated that the effects of genotype (G), environment (E) and their interaction (G × E) were statistically significant ($\chi 2 > 0.001$) for all traits under study. Highly significant differences were observed among traits and environments (Figures S9 to S13; Table 1). On average, PH ranged from 38.27 cm (Env4) to 103.45 cm (Env1). SY and SW data ranged from 670.23 kg ha^{-1} (Env4) to 3329.00 kg ha^{-1} (Env5) and 11.96 g (Env4) to 15.50 g (Env7), respectively. As expected, DTF and DTM varied widely, ranging from 30 (Env2) to 47 (Env3) days, and 88 (Env2) to 133 (Env5) days, respectively (Table S6). The high phenotypic variability was confirmed by analysis of deviance, which revealed that all traits were severely influenced by environmental factors, showing significant G × E interaction (Table 1). Over the eight environments, SY was moderately heritable with a value of 56%, whereas SW, DTM, PH and DTF showed high heritabilities: 81.7%, 91.7%, 93.4% and 94.6%, respectively.

Analysis of phenotypic correlation was performed by environment since residual heterogeneity was observed among the environments and the G × E interaction was significant for all traits. In most of the environments, significant and positive phenotypic correlations were observed between SY and SW, with correlation coefficients ranging from 0.15 (Env2) to 0.58 (Env5). SY and SW showed different patterns of phenotypic correlation with DTF and DTM across environments. The same was observed among SY and SW with PH. In most of the environments, PH and SW showed negative significant phenotypic correlations. However, PH, DTF and DTM were highly positively correlated traits, with correlation coefficients ranging from 0.13 for PH and DTF at Env7 to 0.84 for DTM and DTF at Env4 (Table S7).

3.2. Genome wide association across environments and traits

The results based on Bayesian information criterion (BIC) consistently showed a better fit for the Q + K model over either Q or K alone (Table S8). In total, 33, 29, 57, 72 and 40 linkage disequilibrium blocks were significantly associated with SY, SW, PH, DTM and DTF, respectively (Tables 2 to 6).

Table 1- Summary of analysis of deviance for five traits assessed over combined analysis involving 141 soybean cultivars tested at eight environments of Brazil

Effect	SY		SW		PH		DTF		DTM	
	Deviance[a]	LRT (χ^2)	Deviance	LRT(χ^2)	Deviance	LRT(χ^2)	Deviance	LRT(χ^2)	Deviance	LRT(χ^2)
Genotype	33275.8	442.1**	8612.4	454.7**	15924.2	808.1**	13736.8	1229.4**	14343.5	1050.0**
G*E	32880.9	47.2**	8397.7	240.0**	15792.1	676.0**	13397.1	889.7**	13787.4	493.9**
Error	-		-		-		-		-	
E	-	174.35**	-	92.63**	-	726.81**	-	351.82**	-	593.77**
Block/E	-	26.95**	-	4.67**	-	6.72**	-	2.82*	-	5.52**
Complete Model	32833.7		8157.7		15116.1		12507.4		13293.5	

[a] Deviance of the fitted model without the corresponding effects.
LRT = Likelihood Ratio Test.
Environment and Block/E evaluated with F-test.
** = Significant by F and Chi-square (6.63) test at 1%.

The haplotypes blocks explained considerable phenotypic variation: 17.6% to 96.8%, 13.6% to 33.2%, 45.2% to 99.1%, 13.8% to 59.9% and 12.9% to 42.7% for SY, SW, PH, DTM and DTF, respectively (Table 2 to 6).

For SY, 33 haplotype blocks were effectively associated across environments. These haplotypes were identified on chromosomes 5, 11, 12, 15 and 19, and showed uncharacterized gene annotation or were located in intergenic regions. On the other hand, some haplotypes (identified in the environments Env3 and Env6) were associated with the genomic region that encodes DNA-binding RHL1, beta-fructofuranosidase and kinesin-related protein (Table 2).

The haplotype region Gm12_Hap12 encompasses a genomic region of 420 kb and contain the satt568 and satt442 markers, which are related to the seed protein 28-2 and 28-3 QTLs, respectively (Liang et al., 2010; Yang et al., 2011) (Figure 1). Interestingly, QTLs related to reproductive stage and pod maturity were located near this haplotype. The SSR marker satt192, related to seed glycitein 9-7, was found within Gm12_Hap12. The SNPs located at this chromosomal location were confirmed as an exclusive haplotype region because they were associated with seed yield at Env5, which correspond to Cascavel in the growing season 2014/15. Additionally, two annotated genes were identified: Glyma12g075600, classified as a double-stranded RNA-binding protein 2-like, which encodes a ribonuclease III protein, and Glyma12g075700, which encodes a senescence regulator protein in soybean. In Cascavel (Env5), the haplotypes Gm12_Hap12a and Gm12_Hap12b were significantly different than haplotype Gm12_Hap12c. On average, Gm12_Hap12a and Gm12_Hap12b produced 3354.1 kg ha^{-1} and 3509.0 kg ha^{-1}, while haplotype Gm12_Hap12c yielded 2323.0 kg ha^{-1}, 31% and 34% lower than the haplotypes Gm12_Hap12a and Gm12_Hap12b, respectively. These haplotypes were well distributed in our association mapping panel (TAAT 32%, TAAC 45% and CGGT 23%) (Table 2).

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

Env	Position (bp)			SN	Hap_ID	HapA	HF	SY[a]	R^2 (%)	Nearby Genes/ QTLs
	Chr	Start	End							
Env3	9	38523430	38906660	3	Gm9_Hap22a	CCC	29	2126.2a	21.3	DNA-binding protein RHL1-like
					Gm9_Hap22b	CTC	3	1861.8ab		
					Gm9_Hap22c	TTC	61	1761.7b		
					Gm9_Hap22d	TTT	20	1717.9b		
Env5	12	5622210	6052289	4	Gm12_Hap12a	TAAC	55	3509.0a	41.4	uncharacterized LOC102667945
					Gm12_Hap12b	TAAT	37	3354.1a		
					Gm12_Hap12c	CGGT	28	2323.0b		
Env6	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	34	1815.1a	96.8	Beta-fructofuranosidase insoluble isoenzyme 1-like
					Gm19_Hap42b	GCCGGG	88	1219.3a		
					Gm19_Hap42c	ACCGGG	2	374.2b		
					Gm19_Hap42d	AATGAA	-	-		
Env6	10	3962673	4360182	6	Gm10_Hap8a	TATxTA	16	1999.6a	17.6	uncharacterized LOC100499780
					Gm10_Hap8b	CCGCTA	8	1522.3b		
					Gm10_Hap8c	CCGCCG	30	1516.2bc		
					Gm10_Hap8d	TCTxTA	34	1227.9bcd		
					Gm10_Hap8e	CCGCCA	22	1162.2bcd		
					Gm10_Hap8f	TCTCTA	-	-		
					Gm10_Hap8g	TATCTA	-	-		
					Gm10_Hap8h	TCGCTA	3	.		

Table 2, cont.

	Env6	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	1848.2a	50.6	Intergenic
						Gm19_Hap43b	ACG	2	1113.0b		
						Gm19_Hap43c	GCG	89	1112.9b		
						Gm19_Hap43d	GTA	2	.		
	Env6	11	4462645	4806173	5	Gm11_Hap11a	CCxAA	31	1699.9a	45.6	Probable 125 kDa kinesin-related protein-like
						Gm11_Hap11b	TATCA	6	1548.8ab		
						Gm11_Hap11c	CCTAC	21	1144.4bc		
						Gm11_Hap11d	TATAA	10	1060.8bc		
	Env8	5	5621714	5794460	3	Gm5_Hap7a	CAC	18	2027.8a	18.7	uncharacterized LOC100818074
						Gm5_Hap7b	CGT	21	1956.5a		
						Gm5_Hap7c	TAT	78	1743.5a		
	Env8	15	5621714	5794460	3	Gm15_Hap11a	TCC	7	2024.1a	30.1	uncharacterized LOC100785341
						Gm15_Hap11b	CCC	92	1898.9ab		
						Gm15_Hap11c	TTA	27	1510.8b		

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

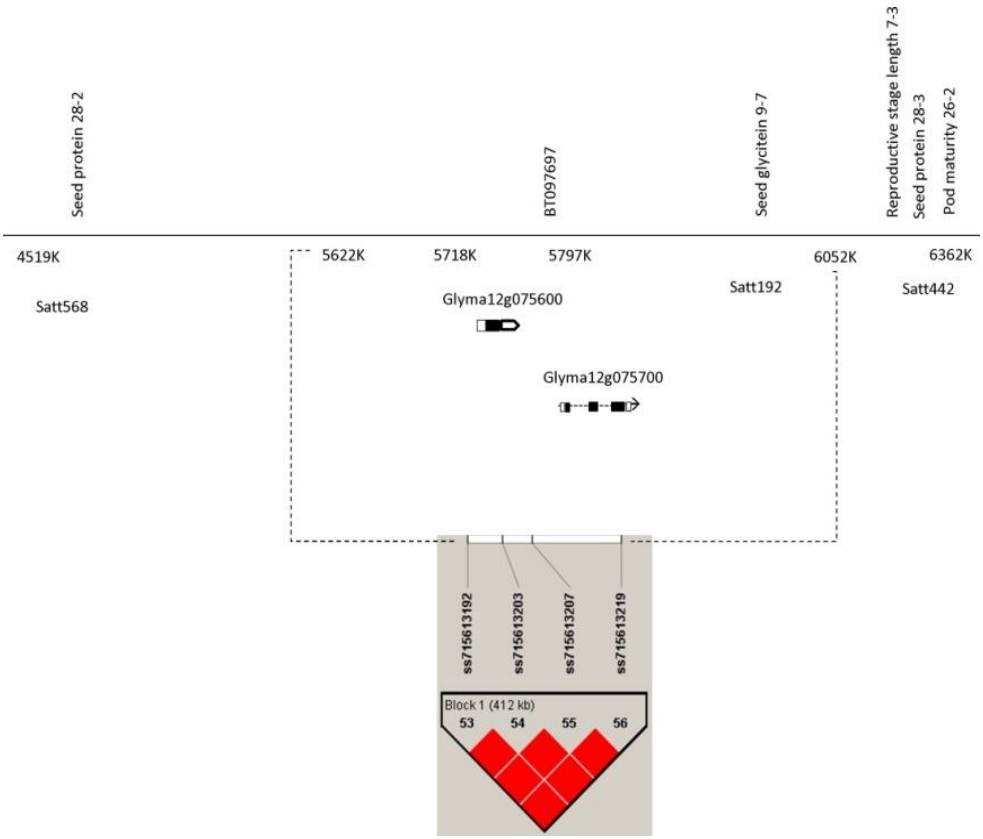


Figure 1 - Candidate region for majors-effect loci: ss715613192 ss715613203, ss715613207 and ss715613219 on Gm12_Hap12 associated with SY, DTF and DTM in soybean. Bottom panel depicts a haplotype region of 412 kb associated with SY and SW (Intensity of red color indicate de r^2, and higher intensity means higher r^2).

For SW, 29 haplotype blocks were effectively associated across environments and chromosomes (Table 3). Gm13_Hap41 was associated with one QTL related to SW, seed weight 40-1 (Rossi et al., 2013), and two others QTLs: Pod maturity 20-1 and Lodging 27-6 (Li et al., 2008; Rossi et al., 2013). Three annotated genes inside this haplotype region were identified: Glyma13g205300, Glyma13g207600 and Glyma13g207900, which encode an unknown protein, a nuclear transcription factor Y (subunit Gamma), and a dihydroxy-acid dehydratase, respectively (Soybase, 2016). Additionally, the haplotype Gm13_Hap42 is associated with SW and contains the annotated gene Glyma13g209500, which encodes a 60S ribosomal protein (Figure 2).

Most of the SNPs effectively associated with PH across environments were located on chromosome 19, including haplotype regions Gm19_Hap42 and

Gm19_Hap43. These haplotypes were consistent across all environments. Gm19_Hap42 is a region containing the Determinate stem 1 gene (Dt1 or GmTFL1) (Cober et al., 2000), found 18.6 kb upstream of the peak SNP ss715635425, which has been previously associated with PH and days to maturity in soybean (Zhang et al., 2015; Contreras-Soto et al., 2017). In addition, other yield QTLs have previously been identified in this region, including seed yield 11-6, and plant height 13-8 and 4-2 (Lee et al., 1996; Specht et al., 2001) (Table 4; Figure 3). Therefore, this QTL region should be considered as a relevant QTL responsible for PH.

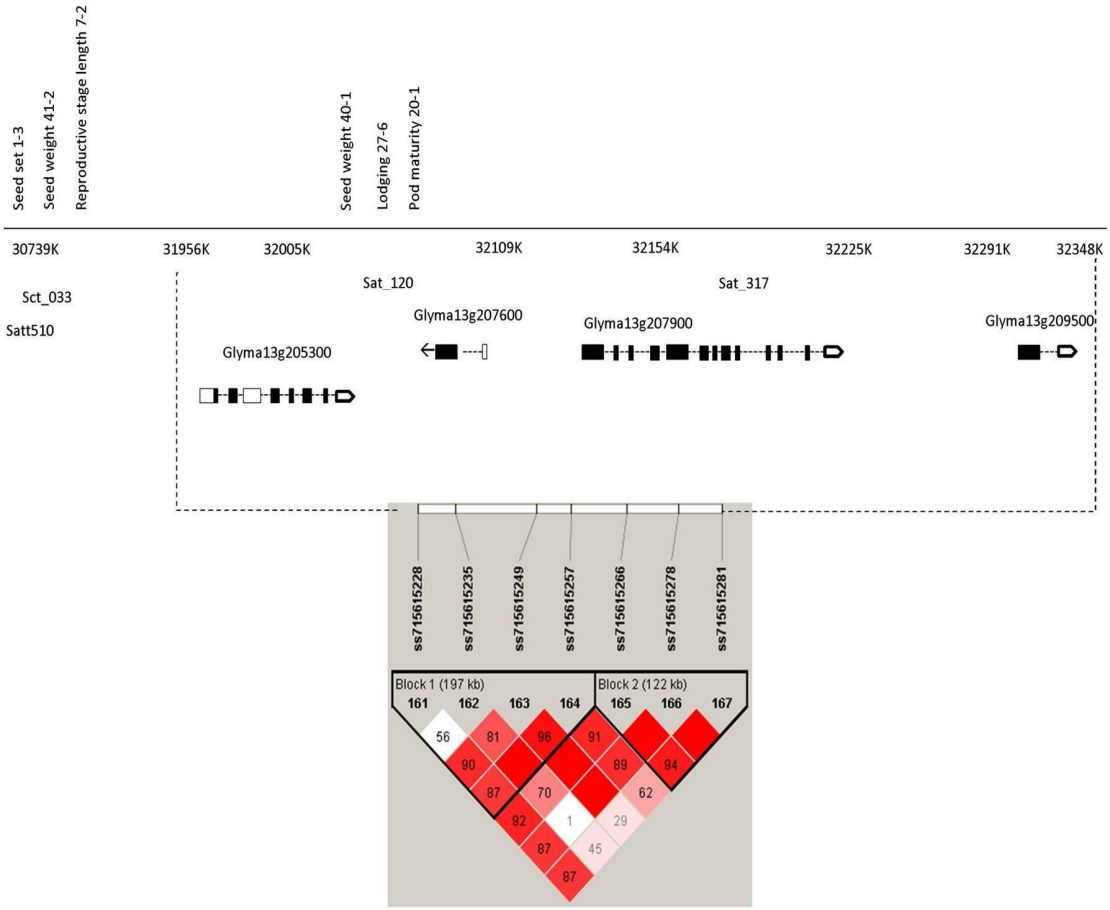


Figure 2 - Candidate region for majors-effect loci: ss715615228, ss715615235, ss715615249 and ss715615257 located on haplotype Gm13_Hap41 and loci ss715615266, ss715615278 and ss715615281 located on haplotype Gm13_Hap42. Gm13_Hap41 was associated with SW, lodging and pod maturity in soybean. Bottom panel depicts haplotypes regions of 197 and 122 kb associated with mention traits (Intensity of red color indicate de r^2, and higher intensity means higher r^2).

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

Env	Position (pb)			SN	Hap_ID	HapA	HF	SW[a]	R^2 (%)	Nearby Genes/ QTLs
	Chr	Start	End							
Env2	11	4875880	4971452	2	Gm11_Hap12a	CT	19	14.0a	27.6	Probable Xaa-Pro aminopeptidase P-like
					Gm11_Hap12b	TC	61	12.5b		
					Gm11_Hap12c	CC	14	11.9b		
					Gm11_Hap12d	TT	22	11.4b		
Env2	11	5074720	5248257	2	Gm11_Hap13a	AA	65	11.9a	13.6	Syntaxin-112 like
					Gm11_Hap13b	GA	18	12.1a		
Env3	11	5074720	5248257	2	Gm11_Hap12a	GA	18	13.4a	15.7	Syntaxin-112 like
					Gm11_Hap12b	AA	65	12.9a		
Env6	13	32225680	32347696	3	Gm13_Hap41a	GAC	51	13.0a	33.2	Glyma13g205300 Glyma13g207600 Glyma13g207900
					Gm13_Hap41b	GGT	31	12.5b		
					Gm13_Hap41c	AAT	28	11.1b		
Env6	13	31956416	32154461	4	Gm13_Hap42a	GTAG	7	14.0a	22.3	Glyma13g209500
					Gm13_Hap42b	GCGG	23	12.7a		
					Gm13_Hap42c	GTGG	44	12.7a		
					Gm13_Hap42d	GTAA	6	12.6a		
					Gm13_Hap42e	ATAA	25	10.9ab		
					Gm13_Hap42f	ACAG	1	10.3ab		
Env7	9	42458021	42790738	4	Gm9_Hap27a	TTTA	18	16.1a	20.2	Auxin-responsive protein IAA8-like
					Gm9_Hap27b	GCTA	76	15.8a		
					Gm9_Hap27c	GCCG	31	14.3b		
Env8	2	8544380	8819494	4	Gm2_Hap22a	AATG	4	16.3a	17.9	Auxilin-like protein 1-like
					Gm2_Hap22b	ACCA	24	15.9ab		
					Gm2_Hap22c	AATA	15	14.7bc		
					Gm2_Hap22d	GCTG	74	14.5bc		
Env8	11	5303401	5800217	4	Gm11_Hap14a	CATC	21	16.4a	23.1	Intergenic
					Gm11_Hap14b	TCTC	45	14.5b		
					Gm11_Hap14c	TATT	24	14.5b		
					Gm11_Hap14d	TCCC	12	14.2b		
					Gm11_Hap14e	TCTT	7	14.1b		

Env: Environment; Chr: Chromosome; SN: Number of SNPs by haplotype; Hap_ID: Haplotype ID; HapA: Allelic haplotypes; HF: Haplotype frequency; SW: mean for 100-seed weight (g/100seed) of haplotypes at each environment.

[a] = Different letter means statistical differences.

For PH, interesting or discriminant haplotypes were located in our association mapping panel, i.e., the haplotype Gm19_Hap43c (GCG), which was associated in most of the environments and showed significant differences with the haplotype responsible for tallest plants (Gm19_Hap43d) in Env3 and Env5. On average, soybean plants with this haplotype showed heights of 93.5 and 84.4 cm of height in Env3 and Env5, respectively, and represented 72% of the total panel (Table 4). However, in Env6, the haplotype Gm19_Hap43a (ATA) was significantly different than the haplotype responsible for smaller plants (Gm19_Hap43c), and consequently produced higher seed yield plants with significant differences among the others haplotypes (Gm19_Hap43a =1848.2 kg ha^{-1}, yielding 28% more than the mean of Env6) (Table 2 and 4).

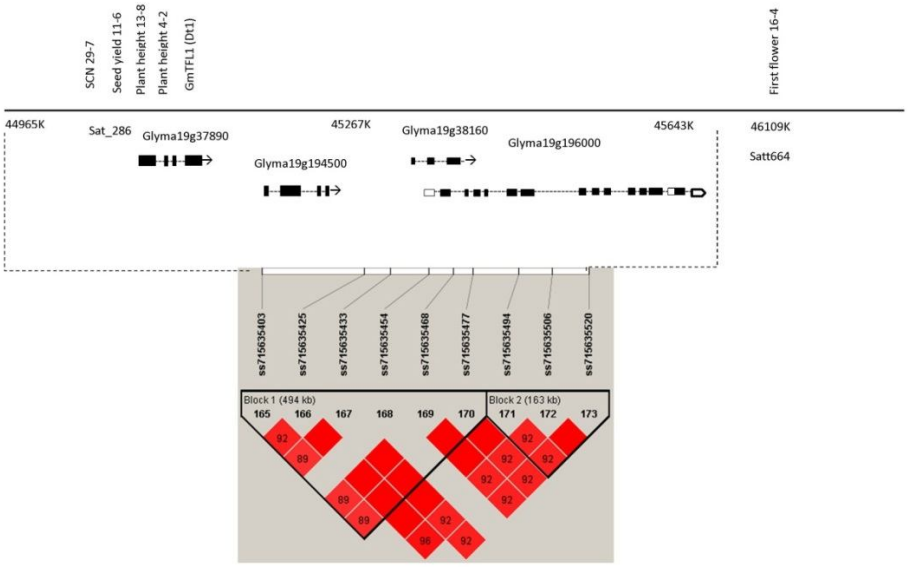


Figure 3 - Candidate region for majors-effect loci: ss715635403, ss715635425, ss715635433, ss715635454 and ss715635468 located on Gm19_Hap42 and loci ss715635494, ss715635506 and ss715635520 located on Gm19_Hap43. Gm19_Hap42 was associated with PH, SY and SCN in soybean. Bottom panel depicts haplotypes regions of 494 (Gm19_Hap42) and 163 kb (Gm19_Hap43) associated with mention traits (Intensity of red color indicate de r^2, and higher intensity means higher r^2).

The haplotype Gm19_Hap42a should differentiate indeterminate growth type in soybean cultivars, whereas Gm19_Hap42b should differentiate determinate soybean cultivars. Gm19_Hap42b showed significant differences with the haplotype responsible for the tallest plants (Gm19_Hap42a) at environments Env1, Env3, Env5,

Env6, Env7 and Env8 (Table 4). Interestingly, in Env6 for SY, this haplotype was not significantly different from the plants that yielded more (Table 2).

For DTM, seventy-two haplotypes were associated across six environments. Of these, forty-one were located in intergenic regions and did not contain genes related to DTM. The other haplotypes were located in genomic regions that encode uncharacterized genes or related proteins (Table 5). The haplotype Gm20_Hap32 is a genomic region that encodes a splicing factor U2AF-associated protein 2-like (LOC100789709 gene) (Table 5). In this candidate region, the following yield loci have previously been associated: seed yield 12-3 and 15-15, plant height 14-1 and 26-15, and seed weight 36-5 (Yuan et al., 2002; Kabelka et al., 2004; Sun et al., 2006; Han et al., 2012). In addition, six annotated genes (Glyma20g218800, Glyma20g220600, Glyma20g221200, Glyma20g222500, Glyma20g222000 and Glyma20g224000) were located within this haplotype region. Annotation of these six genes revealed that they are classified as splicing factor U2AF-associated protein, beta catenin-related armadillo repeat-containing, GDSL Esterase/Lipase, serine/threonine-protein phosphatase PP1 isozyme 2-related, At-hook motif nuclear-localized protein 19-related and Trihelix transcription factor GTL2, respectively (Figure 4).

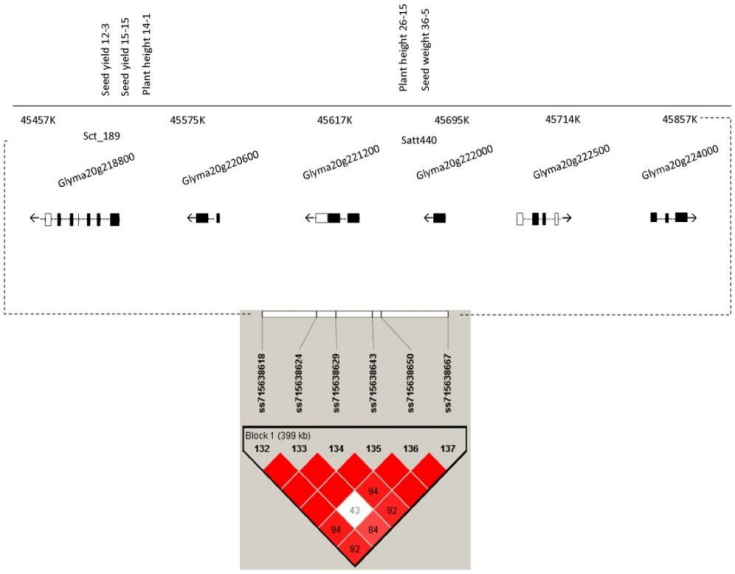


Figure 4 - Candidate region for majors-effect loci: ss715638618, ss715638624, ss715638629, ss715638643, ss715638650 and ss715638667 located on Gm20_Hap32 and associated with DTM. Additionally, QTLs for SY, SW and PH were identified. Bottom panel depicts haplotype region of 399kb associated with mention traits (Intensity of red color indicate de r^2, and higher intensity means higher r^2).

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

Env	Position (bp)			NS	Hap_ID	HapA	HF	PH [a]	R^2 (%)	Nearby Genes/ QTLs
	Chr	Start	End							
Env1	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	121.3a	52.0	Intergenic
					Gm19_Hap43a	ATA	31	115.4a		
					Gm19_Hap43c	GCG	89	98.9a		
					Gm19_Hap43b	ACG	2	90.0a		
Env1	19	44965128	45370594	6	Gm19_Hap42a	AATxAA	31	114.7a	99.1	Sd yld 11-6 Pl ht 4-2 Pl ht 13-8 Dt1 gene (GmTFL1) Sat_286*
					Gm19_Hap42b	GCCGGG	88	99.2b		
					Gm19_Hap42c	ACCGGG	2	83.8b		
Env1	13	36964799	37050736	3	Gm13_Hap53a	GGA	8	120.6a	23.9	uncharacterized LOC102670348
					Gm13_Hap53b	GGG	45	102.4ab		
					Gm13_Hap53c	AAA	53	100.9b		
					Gm13_Hap53d	GAA	4	91.3bc		
Env2	19	45478438	45643073	3	Gm19_Hap43d	GTA	2	55.0a	63.4	Intergenic
					Gm19_Hap43a	ATA	31	53.9a		
					Gm19_Hap43c	GCG	89	46.7a		
					Gm19_Hap43b	ACG	2	43.8a		
Env2	14	8027761	8527621	6	Gm14_Hap21a	CGGGTA	3	63.8a	46.1	Intergenic
					Gm14_Hap21b	CGGGGA	28	54.3a		
					Gm14_Hap21c	TTTAGA	16	49.5ab		
					Gm14_Hap21d	CGTATA	7	48.2ab		
					Gm14_Hap21e	TTTATA	39	47.9b		
					Gm14_Hap21f	TTTAGG	13	44.6b		
					Gm14_Hap21g	CGTAGA	1	.		
Env3	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	109.3a	51.6	Intergenic
					Gm19_Hap43d	GTA	2	106.3a		
					Gm19_Hap43c	GCG	89	93.5b		
					Gm19_Hap43b	ACG	2	83.8b		

Table 4, cont.

					Gm19_Hap42a	AATxAA	31	108.8a		
Env3	19	44965128	45370594	6	Gm19_Hap42b	GCCGGG	88	93.7b	99.1	-*
					Gm19_Hap42c	ACCGGG	2	75.0b		
					Gm19_Hap42a	AATxAA	31	103.8a		
Env5	19	44965128	45370594	6	Gm19_Hap42b	GCCGGG	88	84.6b	99.1	-*
					Gm19_Hap42c	ACCGGG	2	63.8b		
					Gm19_Hap43d	GTA	2	111.3a		
Env5	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	105.3a	55.5	Intergenic
					Gm19_Hap43c	GCG	89	84.4b		
					Gm19_Hap43b	ACG	2	82.5b		
					Gm19_Hap42a	AATxAA	31	94.8a		
Env6	19	44965128	45370594	6	Gm19_Hap42b	GCCGGG	88	63.2b	99.4	-*
					Gm19_Hap42c	ACCGGG	2	32.6b		
					Gm19_Hap43a	ATA	31	94.2a		
Env6	19	45478438	45643073	3	Gm19_Hap43d	ACG	2	69.0ab	79.0	Intergenic
					Gm19_Hap43c	GCG	89	63.2b		
					Gm19_Hap43d	GTA	2	.		
					Gm19_Hap43d	GTA	2	66.3a		
Env7	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	55.8b	64.7	Intergenic
					Gm19_Hap43c	GCG	89	41.4b		
					Gm19_Hap43b	ACG	2	33.8bc		
					Gm19_Hap42a	AATxAA	31	54.7a		
Env7	19	44965128	45370594	6	Gm19_Hap42c	ACCGGG	2	43.8ab	98.2	_*
					Gm19_Hap42b	GCCGGG	88	41.2b		
					Gm19_Hap43d	GTA	2	68.9a		
Env8	19	45478438	45643073	3	Gm19_Hap43a	ATA	31	68.1a	45.2	Intergenic
					Gm19_Hap43c	GCG	89	50.5ab		
					Gm19_Hap43b	ACG	2	44.5ab		
					Gm19_Hap42a	AATxAA	31	67.1a		
Env8	19	44965128	45370594	6	Gm19_Hap42c	ACCGGG	2	57.5ab	69.1	-*
					Gm19_Hap42b	GCCGGG	88	50.4b		

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

For DTF, forty haplotypes were associated across six environments. Most of these were located in intergenic regions of different chromosomes and showed no relationship with genes or markers. On the other hand, three candidate genomic regions that encode cysteine synthase-like, micronuclear linker histone polyprotein-like and APO protein 3 mitochondrial-like were associated with DTF.

The haplotype Gm12_Hap12 was associated with DTF in environments Env1 and Env3, and, interestingly, the same haplotype was associated with SY (Table 2 and 6). Specifically, for DTF and SY, the haplotypes Gm12_Hap12a (TAAC) and Gm12_Hap12b (TAAT) showed significant differences with Gm12_Hap12c (54 and 55 days, respectively). In fact, these haplotypes showed the lowest days to flowering (precocity) (46 and 47 days, and 44 and 46 days in Env1 and Env3, respectively) and the highest yielding plants when compared with Gm12_Hap12c (Table 6).

3.3. Phenotypic variation and correlation between traits

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

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

Env	Position (bp)			NS	Hap_ID	HapA	HF	DTM[a]	R^2 (%)	Nearby Genes/ QTLs
	Chr	Start	End							
Env1	20	45458003	45857761	6	Gm20_Hap32a	GGGGGC	1	150.5a	59.9	LOC100789709 splicing factor U2AF-associated protein 2-like
					Gm20_Hap32b	GGGGAA	1	140.5b		
					Gm20_Hap32c	GAGGGA	8	132.9b		
					Gm20_Hap32d	GGGGGA	22	126.8b		
					Gm20_Hap32e	GAGGGC	67	126.2b		
					Gm20_Hap32f	AAAAAA	19	116.3bc		
Env2	5	5621714	5794460	3	Gm5_Hap7a	CGT	21	98.7a	17.4	Intergenic
					Gm5_Hap7b	TAT	78	87.4b		
					Gm5_Hap7c	CAC	18	87.1b		
Env2	9	6027763	6079042	2	Gm9_Hap13a	CA	14	94.7a	15.5	Intergenic
					Gm9_Hap13b	TA	53	87.6b		
					Gm9_Hap13c	CC	50	87.0b		
Env2	9	43818290	44104810	3	Gm9_Hap30a	CCA	10	89.3a	18.2	Transcription initiation factor TFIID subunit 1-like
					Gm9_Hap30b	CTA	59	87.5a		
					Gm9_Hap30c	CCG	16	87.4a		
					Gm9_Hap30d	TTA	33	87.2a		
Env2	1	49910518	50206347	5	Gm1_Hap17a	TxATA	15	95.5a	53.4	Intergenic
					Gm1_Hap17b	GGGGC	8	88.9ab		
					Gm1_Hap17c	TGGGC	83	87.4ab		
Env2	4	45298627	45435298	2	Gm4_Hap25a	CC	47	89.2a	21.3	Intergenic
					Gm4_Hap25b	CT	16	89.1a		
					Gm4_Hap25c	TT	47	84.8b		
Env2	11	7368580	7405714	2	Gm11_Hap18a	TA	37	89.3a	14.1	Intergenic
					Gm11_Hap18b	Cx	94	87.2a		
Env2	5	2440984	2911445	6	Gm5_Hap2a	GGAAAA	3	91.3a	19.8	LOC100813996 transportin-3-like
					Gm5_Hap2b	GGGGAC	41	89.6a		
					Gm5_Hap2c	GGGAGC	11	87.8a		
					Gm5_Hap2d	TAAAAA	57	87.3a		

Table 5, cont.

					Gm5_Hap2e	GAGAGC	1	.	
					Gm2_Hap3a	TAAT	17	95.8a	
					Gm2_Hap3b	CGAC	13	89.2a	
Env2	2	831795	1033638	4	Gm2_Hap3c	TGAT	31	88.2a	19.2	ATG8i protein
					Gm2_Hap3d	CGCT	19	87.9a	
					Gm2_Hap3e	CGCC	19	86.9a	
					Gm2_Hap3f	CGAT	7	85.4a	
					Gm9_Hap28a	TTA	14	91.9a	
					Gm9_Hap28b	CCG	90	87.8ab	
Env2	9	43003730	43315338	3	Gm9_Hap28c	CTA	3	87.6ab	18.2	uncharacterized LOC100793859
					Gm9_Hap28d	TTG	6	87.6ab	
					Gm9_Hap28e	TCG	11	84.6b	
					Gm7_Hap33a	ATGTT	22	127.7a	
Env3	7	16625092	16979586	5	Gm7_Hap33b	GCACT	45	123.7a	20.1	Intergenic
					Gm7_Hap33c	GCACC	54	122.5a	
					Gm9_Hap26a	GxTTCTA	55	126.2a	
Env3	9	41903227	42370093	7	Gm9_Hap26b	AATTTTA	29	124.9ab	31.1	Intergenic
					Gm9_Hap26c	GAxGCCC	11	120.3ab	
					Gm9_Hap26d	GAxGCTC	15	113.8b	
					Gm2_Hap48a	CAAT	14	141.1a	
Env5	2	40565506	40813466	4	Gm2_Hap48b	CGGC	88	136.1b	21.8	uncharacterized LOC100819417
					Gm2_Hap48c	CGAT	5	127.9bc	
					Gm2_Hap48d	AAAT	13	121.8bc	
					Gm2_Hap33a	AAC	5	123.8a	
					Gm2_Hap33b	ACC	10	120.6a	
Env6	2	13674975	14161558	3	Gm2_Hap33c	ACA	12	119.9a	21.8	Intergenic
					Gm2_Hap33d	GCA	55	119.9a	
					Gm2_Hap33e	GAC	1	119.0a	
					Gm2_Hap33f	AAA	22	116.3a	
Env7	16	30267608	30519426	5	Gm16_Hap26a	GGGCG	111	106.6a	34.1	Intergenic
					Gm16_Hap26b	AATAA	18	97.7b	

Table 5, cont.

					Gm9_Hap19a	AC	32	111.9a	
Env7	9	32388671	32695242	2	Gm9_Hap19b	GC	47	107.5b	12.7	Intergenic
					Gm9_Hap19c	AT	35	99.1c	
					Gm19_Hap10a	GG	61	109.8a	
Env7	19	7322454	7358532	2	Gm19_Hap10b	AG	11	108.0a	23.9	Intergenic
					Gm19_Hap10c	AA	48	98.9b	
					Gm19_Hap11a	GCT	10	109.8a	
					Gm19_Hap11b	GCC	61	109.7a		
Env7	19	8115198	8436529	3	Gm19_Hap11c	TTC	18	102.0b	27.5	Intergenic
					Gm19_Hap11d	TTT	25	97.3b	
					Gm4_Hap31a	CA	2	109.5a	
Env7	4	47740685	48222393	2	Gm4_Hap31b	CG	109	107.6a	13.8	Intergenic
					Gm4_Hap31c	TA	16	98.4b	

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

3.4. Haplotype by environment interaction

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

On the other hand, for plant height, strong and consistent genomic regions within haplotypes across environments were identified (i.e., Gm19_Hap42; Gm19_Hap43). For example, in Cascavel environments, the same haplotype region (Gm19_Hap42) was associated with plant height in the 2012/13, 2013/14 and 2014/15 growing seasons (Env1, Env3 and Env5), and explained most phenotypic variation (99.14%). Specifically, the haplotypes Gm19_Hap42a (AATxAA) and Gm19_Hap42b (GCCGGG) may help in marker-assisted selection of indeterminate and determinate growth habit soybean cultivars, respectively. QTLs controlling plant height are spread over all 20 chromosomes (Soybase, 2016); however, this QTL region could be considered a relevant QTL responsible for PH (Contreras-Soto et al., 2017). In fact, Zhang et al. (2015) previously reported this region as associated with PH and DTM in soybean. In soybean, stem growth habit is regulated by an epistatic interaction between two genes, Dt1 and Dt2 (Bernard, 1972). Dt1 maintains the indeterminate growth habit (dt1dt1 plants are fully determinate), whereas Dt2, in the presence of Dt1, produces semideterminate plants. Dt1 is incompletely dominant over dt1, while Dt2 is completely dominant over dt2. Additionally, our study reported a seed yield QTL in this region. As plant height is one of the major factors determining yield potential in soybean, Gm19_Hap42 (with its large effect on plant height) may also affect soybean yield substantially, as previously reported by Zhang et al. (2015). In addition, the results from Kato et al. (2015) suggest that the indeterminate growth habit is an advantageous characteristic in breeding for high yield of early maturing soybean varieties; however, from the present study, it is clear that the application of the preferred haplotype region needs is complicated, as it may also affect maturity dates, and validation of this haplotype should be improved.

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

Env	Chr	Position (bp) Start	End	NS	Hap_ID	HapA	HF	DTF[a]	R^2 (%)	Nearby Genes/ QTLs
Env1	12	5622210	6052289	4	Gm12_Hap42c	CGGT	28	53.9a	34.6	uncharacterized LOC102667945*
					Gm12_Hap42a	TAAC	55	45.6b		
					Gm12_Hap42b	TAAT	37	43.7b		
Env3	12	5622210	6052289	4	Gm12_Hap42c	CGGT	28	54.9a	41.9	-*
					Gm12_Hap42a	TAAC	55	46.7b		
					Gm12_Hap42b	TAAT	37	45.8b		
Env3	17	8794927	9008173	4	Gm17_Hap10a	GCCG	67	51.6a	38.7	Intergenic
					Gm17_Hap10b	AATA	48	42.1b		
Env3	15	49446994	49521249	2	Gm15_Hap45a	CC	61	52.8a	18.5	LOC100804065 cysteine synthase-like
					Gm15_Hap45b	AC	5	49.0ab		
					Gm15_Hap45c	CT	2	45.5ab		
					Gm15_Hap45d	AT	61	43.8b		
Env4	12	14306367	14775930	5	Gm12_Hap21a	TTCAT	40	43.2a	42.7	Intergenic
					Gm12_Hap21b	CCTGG	79	39.5b		
Env4	9	6155810	6470091	4	Gm9_Hap14a	GGCA	19	46.5a	26.5	Intergenic
					Gm9_Hap14b	GACA	39	40.0a		
					Gm9_Hap14c	AATG	43	38.9a		
					Gm9_Hap14d	AACA	7	38.0a		
					Gm9_Hap14e	AATA	7	35.6b		
Env5	6	50711282	50936449	5	Gm6_Hap52a	TTGCG	8	56.6a	26.1	Intergenic
					Gm6_Hap52b	CTGCG	20	49.8ab		
					Gm6_Hap52c	TGGCG	10	48.6ab		
					Gm6_Hap52d	CGGCG	39	48.6ab		
					Gm6_Hap52e	TTGTA	6	47.4ab		
					Gm6_Hap52f	TTATA	26	40.1b		
Env5	12	38680709	38970900	2	Gm12_Hap35a	AG	5	62.2a	12.9	LOC102660802 micronuclear linker histone polyprotein-like
					Gm12_Hap35b	AT	65	48.4b		
					Gm12_Hap35c	CG	8	45.9bc		
					Gm12_Hap35d	CT	43	43.9c		

Table 6, cont.

Env5	20	41883051	42297577	4	Gm20_Hap27a Gm20_Hap27b Gm20_Hap27c	GCGG ACGG ATTA	11 19 91	52.7a 50.2a 45.9a	32.9	Intergenic
Env7	2	41787747	42088045	4	Gm2_Hap51a Gm2_Hap51b Gm2_Hap51c Gm2_Hap51d	GGCG GATA TGTA TATA	11 44 6 58	42.6a 41.6a 34.0b 33.9b	18.1	APO protein 3, mitochondrial-like**
Env8	2	41787747	42088045	4	Gm2_Hap51a Gm2_Hap51b Gm2_Hap51d Gm2_Hap51c	GGCG GATA TATA TGTA	11 44 58 6	36.2a 34.4a 28.8b 28.6b	19.8	-**

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

3.5. Co-associated haplotype genomic regions among yield and flowering traits

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

The LOC100789709 gene on chromosome 20 (Gm20_Hap32), described as a splicing factor U2AF-associated protein, was related to DTM in soybean. This gene is a homolog of atU2AF in Arabidopsis thaliana. Wang and Brendel (2006) demonstrated that altered expression levels of atU2AF35a or atU2AF35b causes pleiotropic phenotypes in flowering time, leaf morphology, flower, and silique shape in A. thaliana; specifically, pleiotropic phenotypes have been observed in mutants and transgenic lines. Homozygous atU2AF35a T-DNA insertion plants and atU2AF35b transgenic plants showed late flowering under both long and short day conditions. In fact, the altered expression of this gene may also affect days to flowering and maturity in soybean, confirming the haplotype association with this latter trait. Additionally, in this candidate region, some loci controlling grain yield have previously been associated: seed yield 12-3 and 15-15, plant height 14-1 and 26-15,

and seed weight 36-5 (Yuan et al., 2002; Kabelka et al., 2004; Sun et al., 2006; Han et al., 2012). These results suggest that the morphological correlations between yield components and time to flowering and maturity traits are related on a genetic basis, suggesting gene pleiotropy and high rates of linkage disequilibrium (Chen and Lubberstedt, 2010).

On chromosome 12 the haplotype Gm12_Hap12 was significantly associated with SY, DTF and DTM traits in all environments under study. This result may suggest that this region contains a single gene that has pleiotropic effects and is tightly linked with multiple genes. Recker et al. (2014) evaluated multiple environments to show that SY and DTM are positively correlated, while SY was not significantly correlated with DTF. In the present study, variable correlation results were obtained at individual environments, e.g., for SY and DTM: r=-0.65 (at Env5-Cascavel) to r=0.39 and 0.26 (at Env2-Primavera do Leste and Env7-Rio Verde, respectively); For SY and DTF: r=-0.44 (at Env1-Cascavel) to r=0.46 (at Env7-Rio Verde). As such, these results should be interpreted at the environment level considering that these traits exhibit QTL-by-environment interactions. In Cascavel, the haplotype Gm12_Hap12 should be used to improve yield and precocity in the current soybean program. Specifically, the haplotypes Gm12_Hap12a and Gm12_Hap12b showed significant differences from Gm12_Hap12c for DTF and SY. In fact, these haplotypes showed the lowest days to flowering (precocity) (46 and 47 days, and 44 and 46 days, respectively) and the highest yield plants when compared with Gm12_Hap12c. Furthermore, the fine mapping of such regions could help to discern the specific genetic elements controlling these traits. In this case, we used haplotypes to obtain the best performance for each trait and environment.

Finally, the results of this study suggest that the BARCSoySNP6K BeadChip and haplotype-based genome-wide association are valuable sources of information for discovering genomic regions that control quantitative traits in soybean. This research identified useful associated markers that have not been previously reported and that were detected in multiple environments. This will facilitate assessing and validating causal genetic variation of complex quantitative traits and may eventually be used to accelerate the optimization of molecular breeding. However, as with any molecular markers, we emphasize that the identified haplotypes should be validated before large-scale use.

4. REFERENCES

ABDEL-SHAFY, H.; BORTFELDT, R.H.; TETENS, J.; BROCKMANN, G.A. Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. **Genetics Selection Evolution,** 46:35, 2014.

BARRETT, J.C.; FRY, B.; MALLER, J.; DALY, M.J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics,** 21:263-265, 2005.

BERNARD, R.L. Two genes affecting stem termination in soybeans. **Crop Science,** 12: 235-239, 1972.

BRADBURY, P.; ZHANG, Z.; KROON, D.; CASSTEVENS, T.; RAMDOSS, Y.; BUCKLER, E. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics,** 23:2633-2638, 2007.

BRADLEY, D.; CARPENTER, R.; COPSEY, L.; VINCENT, C.; ROTHSTEIN, S.; COEN, E. Control of inflorescence architecture in Antirrhinum. **Nature,** 379:791–797, 1996.

CAPPA, E.P.; EL-KASSABY, Y.A.; GARCIA, M.N.; ACUÑA, C.; BORRALHO, N.M.; GRATTAPAGLIA, D.; MARCUCCI POLTRI, S.N. Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: a case study in *Eucalyptus globulus*. **PLoS ONE,** 8:e81267, 2013.

CHEN, Y.; LUBBERSTEDT, T. Molecular basis of trait correlations. **Trends of Plant Science,** 15: 454-461, 2010.

COBER, E.R.; MADILL, J.; VOLDENG, H.D. Early tall determinate soybean genotype E1E1e3e3e4e4dt1dt1 sets high bottom pods. **Canadian Journal of Plant Sciences,** 80:527-531, 2000.

COBER, E.R.; MORRISON, M.J. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. **Theoretical and Applied Genetics,** 120:1005-1012, 2010.

CONTRERAS-SOTO, R.; MORA, F.; OLIVEIRA, M.A.R.; HIGASHI, W.; SCAPIM, C.A.; SCHUSTER, I. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. **PLoS ONE,** e0171105, 2017.

EMBRAPA. **Tecnologias de produção de soja** – região central do Brasil 2012 e 2013. Londrina: Embrapa Soja, 2011. 261p. (Sistemas de Produção / Embrapa Soja, n.15).

ENDELMAN, J.B.; JANNINK, J.C. Shrinkage estimation of realized relationship matrix. **Genes Genomes Genetics,** 2:1405-1413, 2012.

FEHR, W.R.; CAVINESS, C.E. **Stages of soybean development**. Cooperative Extension Service, Agriculture and Home Economics Experiment Station, Iowa: Ames, 1977. 12p.

FOX, C.M.; CARY, T.R.; NELSON, R.L.; DIERS, D.W. Confirmation of a Seed Yield QTL in Soybean. **Crop Science,** 55:992-998, 2015.

GAO, H.; WILLIAMSON, S.; BUSTAMANTE, C.D. A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. **Genetics,** 176:1635-1651, 2007.

GARNER, C.; SLATKIN, M. On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. **Genetic Epidemiology,** 24: 57-67, 2003.

GREENSPAN, G.; GEIGER, D. Model-based inference of haplotype block variation. **Journal of Computational Biology,** 11:493-504, 2004.

GUZMAN, P.S.; DIERS, B.W.; NEECE, D.J.; ST. MARTIN, S.K.; LEROY, A.R.; GRAU, C.R.; HUGHES, T.J.; NELSON, R.L. QTL associated with yield in three backcross-derived populations of soybean. **Crop Science,** 47:111-122, 2007.

HAN, Y.; LI, D.; ZHU, D.; LI, H.; LI, H.; TENG, W.; LI, W. QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. **Theoretical and Applied Genetics,** 125:671-683, 2012.

HAO, D.; CHENG, H.; YIN, Z.; CUI, S.; ZHANG, D.; WANG, H.; YU, D. Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. **Theoretical and Applied Genetics,** 124:447–458, 2012.

KABELKA, E.A.; DIERS, B.W.; FEHR, W.R.; LEROY, A.R.; BAIANU, I.C.; YOU, T.; NEECE, D.J.; NELSON, R.L. Putative alleles for increased yield from soybean plant introductions. **Crop Science,** 44:784–791, 2004.

KATO, S.; FUJII, K.; YUMOTO, S.; ISHIMOTO, M.; SHIRAIWA, T.; SAYAMA, T.; KIKUCHI, A.; NISHIO, T. Seed yield and its components of indeterminate and determinate lines in recombinant inbred lines of soybean. **Breeding Science,** 65:154-160, 2015.

KIM, K.S.; DIERS, D.W.; HYTEN, D.L.; MIAN, M.A.R.; SHANNON, J.G.; NELSON, R.L. Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. **Theoretical and Applied Genetics,** 125:1353–1369, 2012.

LEE, S.H.; BAILEY, M.A.; MIAN, M.A.R.; CARTER, T.E.; ASHLEY, D.A.; HUSSEY, R.S.; PARROTT, W.A.; BOERMA, H.R. Molecular markers associated with soybean plant height, lodging, and maturity across locations. **Crop Science,** 36:728-735, 1996.

LI, D.; PFEIFFER, T.W.; CORNELIUS, P.L. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. **Crop Science,** 48:571-581, 2008.

LI, W.; ZHENG, D.H.; VAN, K.; LEE, S.H. QTL Mapping for major agronomic traits across two years in soybean (*Glycine max* L. Merr.). **Journal of Crop Science and Biotechnology,** 11:171-190, 2008.

LIANG, H.; YU, Y.; WANG, S.; LIAN, Y.; WANG, T.; WEI, Y.; GONG, P.; LIU, X.Y.; FANG, X.J.; ZHANG, M.C. QTL Mapping of Isoflavone, Oil and Protein Contents in Soybean (*Glycine max* L. Merr.). **Agricultural Sciences in China,** 9:1108-1116, 2010.

LIU, B.; WATANABE, S.; UCHIYAMA, T.; KONG, F.; KANAZAWA, A.; XIA, Z.; NAGAMATSU, A.; ARAI, M.; YAMADA, T.; KITAMURA, K.; MASUTA, C.; HARADA,

K.; ABE, J. The soybean stem growth habit gene Dt1 is an ortholog of Arabidopsis TERMINAL FLOWER1. **Plant Physiology,** 153:198-210, 2010.

LORENZ, A.J.; HAMBLIN, M.T.; JANNINK, J.L. Performance of single nucleotide polymorphisms versus haplotypes for Genome-Wide Association analysis in Barley. **PLoS ONE,** 5:e14079, 2010.

MANSUR, L.M.; ORF, J.H.; CHASE, K.; JARVIK, T.; CREGAN, P.B.; LARK, K.G. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. **Crop Science,** 36: 1327-1336, 1996.

MORA, F.; QUITRAL, Y.A.; MATUS, I.; RUSSELL, J.; WAUGH, R.; DEL POZO, A. SNP-based QTL mapping of fifteen complex traits in barley under rain fed and well-watered conditions by a mixed modeling approach. **Frontiers of Plant Science,** 7:909, 2016.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized Linear Models. **Journal of the Royal Statistical Society,** 135:370-384, 1972.

ORF, J.H.; CHASE, K.; JARVIK, T.; MANSUR, L.M.; CREGAN, P.B.; ADLER, F.R.; LARK, K.G. Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. **Crop Science,** 39:1642-1651, 1999.

PALOMEQUE, L.; LIU, L.J.; LI, W.B.; HEDGES, B.R.; COBER, E.R.; SMID, M.P.; LUKENS, L.; RAJCAN, I. Validation of mega-environment universal and specific QTL associated with seed yield and agronomic traits in soybeans. **Theoretical and Applied Genetics,** 120:997-1003, 2010.

RECKER, J.R.; BURTON, J.W.; CARDINAL, A.; MIRANDA, L. Genetic and Phenotypic Correlations of Quantitative Traits in Two Long-Term, Randomly Mated Soybean Populations. **Crop Science,** 54:939-943, 2014.

ROSSI, M.E.; ORF, J.H.; LIU, L.J.; DONG, Z.; RAJCAN, I. Genetic basis of soybean adaptation to North American vs. Asian mega-environments in two independent populations from Canadian × Chinese crosses. **Theoretical and Applied Genetics,** 126:1809-1823, 2013.

SCHUSTER, I. Marker-assisted selection for quantitative traits. **Crop Breeding and Applied Biotechnology,** S1:50-55, 2011.

SCHWARZ, G. Estimating the dimension of a model. **Annals Statistics,** 6:461-464, 1978.

SEBASTIAN, S.A.; STREIT, L.G.; STEPHENS, P.A.; THOMPSON, J.A.; HEDGES, B.R.; FABRIZIUS, M.A.; SOPER, J.F.; SCHMIDT, D.H.; KALLEM, R.L.; HINDS, M.A.; FENG, L.; HOECK, J.A. Context-specific marker-assisted selection for improved grain yield in elite soybean populations. **Crop Science,** 50:1196–1206, 2010.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE,** 8:e54985, 2013.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Fingerprinting soybean germplasm and its utility in genomic research. **Genes Genomes Genetics,** 5:1999-2006, 2015.

SOYBASE (2016). **USDA-ARS Soybean genetics and genomics database**. USDA, Washington, DC. Disponível em: www.soybase.org/search/qtllist.php. Acesso em: 10, agosto, 2016.

SPECHT, J.E.; CHASE, K.; MACRANDER, M.; GRAEF, G.L.; CHUNG, J.; MARKWELL, J.P.; GERMANN, M.; ORF, J.H.; LARK, K.G. Soybean response to water: a QTL analysis of drought tolerance. **Crop Science,** 4:493–509, 2001.

SPIEGELHALTER, D.J.; BEST, N.G.; CARLIN, B.P.; VAN DER LINDE, A. Bayesian measures of model complexity and fit, (with discussion). **Journal of the Royal Statistical Society Series B,** 64:583-639, 2002.

STICH, B.; MOHRING, J.; PIEPHO, H.P.; HECKENBERGER, M.; BUCKLER, E.S.; MELCHINGER, A.E. Comparison of mixed-model approaches for association mapping. **Genetics,** 178:1745–1754, 2008.

SUN, D.; LI, W.; ZHANG, Z.; CHEN, Q.; NING, Q.; QIU, L.; SUN, G. Quantitative trait loci analysis for the developmental behavior of soybean (*Glycine max* L. Merr) **Theoretical and Applied Genetics,** 112:665-673, 2006.

SUN, S.; KIM, M.Y.; VAN, K.; LEE, Y.W.; LI, B.; LEE, S.H. QTLs for resistance to Phomopsis seed decay are associated with days to maturity in soybean (*Glycine max*). **Theoretical and Applied Genetics,** 126:2029-2038, 2013.

TIAN, Z.; WANG, X.; LEE, R.; LI, Y.; SPECHT, J.E.; NELSON, R.L.; MCCLEAN, P.E.; QIU, L.; MA, J. Artificial selection for determinate growth habit in soybean. **Proceedings of the National Academy of Science USA,** 107:8563-8568, 2010.

VICENTE, D.; SCHUSTER, I.; LAZZARI, F.; PARANZINI, J.P.D.; OLIVEIRA DE, M.A.R.; PRETE, C.E.C. Mapping and validation of molecular markers of genes Dt1 and Dt2 to determine the type of stem growth in soybean. **Acta Scientiarum Agronomy,** 38:61-68, 2016.

WANG, B.B.; BRENDEL, V. Molecular characterization and phylogeny of U2AF35 homologs in plants. **Plant Physiology,** 140:624–636, 2006.

YANG, K.; MOON, J.; JEONG, N.; CHUN, H.; KANG, S.; BACK, K.; JEONG, S. Novel major quantitative trait loci regulating the content of isoflavone in soybean seeds. **Genetics and Genomics,** 33:685-692, 2011.

YU, J.; PRESSOIR, G.; BRIGGS, W.; VROH, B.I.; YAMASAKI, M.; DOEBLEY, J.; MCMULLEN, M.; GAUT, B.; NIELSEN, D.; HOLLAND, J.; KRESOVICH, S.; BUCKLER, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics,** 38:203–208, 2006.

YUAN, J.; NJITI, V.N.; MELSEM, K.; IQBAL, J.J.; TRIWITAYAKORN, K.; KASSEM, M.A.; DAVIS, G.T.; SCHMIDT M.E.; LIGHTFOOT D.A. Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. **Crop Science,** 42:271-277, 2002.

ZHANG, H.; HAO, D.; SITOE, H.M.; YIN, Z.; HU, Z.; ZHANG, G.; YU, D.; SINGH, R. Genetic dissection of the relationship between plant architecture and yield

component traits in soybean (*Glycine max*) by association analysis across multiple environments. **Plant Breeding,** 134:564-572, 2015.

ZHANG, J.; SONG, Q.; CREGAN, P.B.; NELSON, R.L.; WANG, X.; WU, J.; JIANG, G.L. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. **BMC Genomics,** 16:217, 2015.

SUPPLEMENTARY FILES

Table S1 - Maturity group (MG), company origin, and population structure membership group (IC), and bar-plot code of population structure of one hundred sixty nine improved tropical soybean cultivars utilized in genome-wide association study

Code	Variety	MG*	IC*	Company	Bar-plot Code IC
1	BMX APOLLORR	5.5	1	GDM	G1
2	BMX MAGNARR	6.2	1	GDM	G2
3	BRS243RR	6.6	1	Embrapa	G3
4	BRS257	6.4	1	Embrapa	G4
5	BRSMS Lambari	7.3	1	Embrapa	G5
6	BRSMT Crixás	8.5	1	Embrapa	G6
7	CD 214RR	6.7	1	Coodetec	G7
8	CD 249RR STS	6.7	1	Coodetec	G8
9	CD 250RR	5.5	1	Coodetec	G9
10	CD 254RR	8.4	1	Coodetec	G10
11	CD 2800	8.0	1	Coodetec	G11
12	CD 2860	8.6	1	Coodetec	G12
13	EMBRAPA 48	6.5	1	Embrapa	G13
14	FUNDACEP 57RR	6.2	1	Bayer	G14
15	M-SOY 6101	6.1	1	Monsanto	G15
16	P98Y70	8.7	1	Pioneer	G16
17	BRS232	6.5	2	Embrapa	G17
18	BRS246RR	7.3	2	Embrapa	G18
19	BRS283	6.5	2	Embrapa	G19
20	CD204	7.3	2	Coodetec	G20
21	CD 213RR	6.8	2	Coodetec	G21
22	CD 218	7.2	2	Coodetec	G22
23	CD 221	6.4	2	Coodetec	G23
24	CD 224	6.9	2	Coodetec	G24
25	CD 225RR	5.8	2	Coodetec	G25
26	CD 239RR	6.7	2	Coodetec	G26
27	CD 244RR	8.0	2	Coodetec	G27
28	CD 246	8.1	2	Coodetec	G28
29	MG/BR 46 (Conquista)	8.2	2	Embrapa	G29
30	DMario 70i	7.0	2	GDM	G30
31	EMBRAPA 59	6.5	2	Embrapa	G31
32	FUNDACEP 55RR	6.0	2	Bayer	G32
33	FUNDACEP 58RR	6.8	2	Bayer	G33
34	BRSMG Liderança	7.7	2	Embrapa	G34
35	M7211RR	7.2	2	Monsanto	G35
36	M8527RR	8.5	2	Monsanto	G36
37	MERCEDES70A	6.6	2	Unknown**	G37
38	M-SOY 7901	7.9	2	Monsanto	G38
39	NA 4990 RG	4.9	2	Nidera	G39
40	FMT TABARANA	8.7	2	Embrapa	G40
41	TMG4001RR	6.5	2	TMG	G41
42	BMX ATIVARR	5.6	3	GDM	G42
43	BMX ForcaRR	6.2	3	GDM	G43
44	BMX POTENCIARR	6.7	3	GDM	G44
45	BRS185	6.6	3	Embrapa	G45

Table S1, cont.

46	BRS259	7.1	3	Embrapa	G46
47	BRS282	6.5	3	Embrapa	G47
48	BRSMG Renascença	7.9	3	Embrapa	G48
49	CD 229RR	7.3	3	Coodetec	G49
50	CD 230RR	7.6	3	Coodetec	G50
51	BRS CELESTE	8.7	3	Embrapa	G51
52	M9144RR	9.1	3	Monsanto	G52
53	MSOY2001	7.9	3	Monsanto	G53
54	M-SOY 8001	8.0	3	Monsanto	G54
55	OC 14	5.8	3	Embrapa	G55
56	OCEPAR3-PRIMAVERA	6.5	3	Embrapa	G56
57	TMG103RR	8.3	3	TMG	G57
58	TMG7161RR	6.5	3	TMG	G58
59	5G830RR	8.3	4	DowAgroscience	G59
60	A6001RR	6.0	4	Unknown	G60
61	ANTA82	7.4	4	Unknown	G61
62	BMX TURBORR	5.8	4	GDM	G62
63	BRSMT Pintado	8.5	4	Embrapa	G63
64	CD 208	6.9	4	Coodetec	G64
65	CD 228	7.5	4	Coodetec	G65
66	CD 236RR	6.2	4	Coodetec	G66
67	CD 251RR	8.8	4	Coodetec	G67
68	CD 2737RR	7.3	4	Coodetec	G68
69	FUNDACEP 59RR	7.5	4	Bayer	G69
70	M6009RR	6.0	4	Monsanto	G70
71	M6707RR	6.7	4	Monsanto	G71
72	M-SOY 7201	7.2	4	Monsanto	G72
73	NA 5909 RG	5.9	4	Nidera	G73
74	A 7321RG	7.3	4	Nidera	G74
75	5G770RR	7.7	5	DowAgroscience	G75
76	A8000	8.0	5	Unknown	G76
77	BMX TitanRR	5.3	5	GDM	G77
78	CD201	6.7	5	Coodetec	G78
79	CD205	7.8	5	Coodetec	G79
80	CD 217	7.3	5	Coodetec	G80
81	CD 233RR	6.4	5	Coodetec	G81
82	CD 242RR	7.9	5	Coodetec	G82
83	CD 5807	6.0	5	Coodetec	G83
84	FUNDACEP 39	7.1	5	Bayer	G84
85	FUNDACEP 53RR	6.4	5	Bayer	G85
86	Embrapa 1 (IAS 5-RC)	6.4	5	Embrapa	G86
87	FMT MATRINXA	7.9	5	Embrapa	G87
88	TMG115RR	8.6	5	TMG	G88
89	5D660RR	6.6	6	DowAgroscience	G89
90	5D688RR	6.8	6	DowAgroscience	G90
91	A6001	6.0	6	Unknown	G91
92	BRS256RR	7.8	6	Embrapa	G92
93	BRS284	6.4	6	Embrapa	G93
94	Capinópolis (UFV-16)	7.7	6	Embrapa	G94
95	CD 215	5.9	6	Coodetec	G95
96	CD 231RR	7.3	6	Coodetec	G96
97	CD 234RR	8.0	6	Coodetec	G97
98	CD 235RR	6.4	6	Coodetec	G98
99	CD 237RR	7.3	6	Coodetec	G99

Table S1, cont.

100	CD 240RR	6.9	6	Coodetec	G100
101	CD 248RR	6.4	6	Coodetec	G101
102	CD 253	8.7	6	Coodetec	G102
103	CD 2792RR	7.9	6	Coodetec	G103
104	CD/FAPA 220	7.3	6	Coodetec	G104
105	EMGOPA 302	6.4	6	Embrapa	G105
106	Fundacep 38	7.0	6	Bayer	G106
107	MG/BR 48 (Garimpo)	7.8	6	Embrapa	G107
108	IGRA RA 626RR	7.7	6	Igra	G108
109	M7639RR	7.6	6	Monsanto	G109
110	M7908RR	7.9	6	Monsanto	G110
111	SYN3358 RR	6.4	6	Syngenta	G111
112	NK 7059 RR	7.0	6	Syngenta	G112
113	NS 4823RR	4.8	6	Nidera	G113
114	P98Y51	8.5	6	Pioneer	G114
115	SPRING53	5.3	6	Syngenta	G115
116	TMG1161RR	6.7	6	TMG	G116
117	5D690RR	6.9	7	DowAgroscience	G117
118	5D711RR	7.1	7	DowAgroscience	G118
119	BRS184	6.1	7	Embrapa	G119
120	BRS268	7.2	7	Embrapa	G120
121	CAC 1	8.3	7	Embrapa	G121
122	CD 219RR	8.2	7	Coodetec	G122
123	CD 238RR	7.1	7	Coodetec	G123
124	CD 2630RR	6.3	7	Coodetec	G124
125	CD 2840	8.4	7	Coodetec	G125
126	CD 5969	6.4	7	Coodetec	G126
127	EMGOPA 304 (Campeira)	7.3	7	Embrapa	G127
128	Fundacep 33	8.0	7	Bayer	G128
129	Fundacep 56RR	6.5	7	Bayer	G129
130	IGRA RA 628RR	6.4	7	Igra	G130
131	A 6411RG	6.4	7	Nidera	G131
132	P98Y11	8.1	7	Pioneer	G132
133	TMG 1066RR	6.6	7	TMG	G133
134	TMG123RR	6.7	7	TMG	G134
135	VENCEDORA	8.0	7	Embrapa	G135
136	Bragg	6.6	8	Embrapa	G136
137	BRS133	6.6	8	Embrapa	G137
138	BRS230	6.4	8	Embrapa	G138
139	BRS258	7.3	8	Embrapa	G139
140	CD202	6.4	8	Coodetec	G140
141	CD206	6.8	8	Coodetec	G141
142	CD 206RR	6.8	8	Coodetec	G142
143	CD 216	5.5	8	Coodetec	G143
144	CD 226RR	6.6	8	Coodetec	G144
145	CD 232	7.3	8	Coodetec	G145
146	FT-ESTRELA	8.0	8	Embrapa	G146
147	FUNDACEP 61RR	6.0	8	Bayer	G147
148	FUNDACEP 63RR	5.4	8	Bayer	G148
149	IGRA RA 516RR	6.4	8	Igra	G149
150	A 4725RG	4.7	8	Nidera	G150
151	TMG 1067RR	6.7	8	TMG	G151
152	TropicalRR	6.7	8	TMG	G152
153	USP1	6.6	8	Unknown	G153

Table S1, cont.

154	BMX ENERGIARR	5.0	9	GDM	G154
155	BRS213	6.5	9	Embrapa	G155
156	BRS245RR	7.3	9	Embrapa	G156
157	BRS262	7.9	9	Embrapa	G157
158	CD 243RR	8.0	9	Coodetec	G158
159	CD 245RR	8.4	9	Coodetec	G159
160	CD 247RR	8.3	9	Coodetec	G160
161	CD 252	6.4	9	Coodetec	G161
162	CD 2585RR	5.8	9	Coodetec	G162
163	CD 2721RR	7.2	9	Coodetec	G163
164	FT Abyara	7.7	9	Embrapa	G164
165	FT-GUAIRA	6.4	9	Embrapa	G165
166	IGRA RA 518RR	6.0	9	Igra	G166
167	M7578RR	7.5	9	Monsanto	G167
168	R7	7.0	9	Unknown	G168
169	FMT TUCUNARE	8.3	9	Embrapa	G169

Table S2 - Goodness of fit of three different GWAS models for: seed yield, 100-seed weight and plant height in 169 varieties of soybean evaluated in four environments of Brazil. Q represents the model with population structure effect; K represents the model with kinship effect and Q + K represent the model with the joint effects

Environment	Models	Seed Yield		100 Seed-Weight		Plant Height	
		-2 log likelihood	BIC	-2 log likelihood	BIC	-2 log likelihood	BIC
Cascavel	Q	2556.74	2577.26	658.32	678.84	1367.47	1387.99
	K	2611.74	2627.13	638.05	653.44	1386.05	1401.44
	Q + K	2477.44	2503.09	612.58	638.23	1328.89	1354.54
Palotina	Q	2025.32	2045.31	452.73	472.69	1034.00	1053.31
	K	2126.72	2141.71	420.00	434.97	1086.45	1100.94
	Q + K	2010.26	2035.25	397.27	422.22	1018.24	1042.38
Primavera do Leste	Q	2532.65	2553.17	609.47	629.99	1035.64	1055.60
	K	2617.51	2632.89	612.66	628.05	1071.15	1086.12
	Q + K	2496.04	2521.68	579.53	605.18	1012.33	1037.28
Rio Verde	Q	1428.59	1446.93	293.91	312.25	725.34	743.72
	K	1534.24	1548.00	311.55	325.31	760.79	774.57
	Q + K	1411.46	1434.39	288.68	311.60	696.99	719.97

Table S3 - Summary of mixed modeling analyses (Q + K model) for SNPs and haplotypes significantly associated with seed yield evaluated in 169 cultivars of soybean in four environment of southern Brazil

Environment	Marker [a]	SNP [b]	Haplotype Block LD	Chr	Position	$log_{10}(P)$	R^2	Nearby Genes
Cascavel	ss715613203	G/A	12	12	5706745	4.12	11.97	Ribonuclease III satt568; satt442 and satt192
	ss715613104	A/C	- [c]	12	4670638	4.28	11.92	-
	ss715613207	A/G	12	12	5786241	4.25	11.92	-[d]
	ss715613192	T/C	12	12	5610878	3.23	10.17	-[d]
Rio Verde	ss715614920	T/C	36	13	28957669	3.22	9.14	Putative germinal-center associated nuclear protein-like
	ss715593323	A/G	28	6	15032691	3.07	15.83	-

Chr: Chromosome; LD: Linkage disequilibrium;
[a] http://soybase.org/snps/
[b] Significant at $-log(P) > 3$
[c] without haplotype
[d] SNPs were associated with the same previous reported QTLs in **

Table S4 - Summary of mixed modeling analyses (Q + K model) for SNPs and haplotypes significantly associated with 100-seed weight evaluated in 169 cultivars of soybean in four environment of southern Brazil

Environment	Marker [a]	SNP [b]	Haplotype Block LD	Chr	Position	$log_{10}(P)$	R^2	Nearby Genes
Cascavel	ss715592623	A/G	10	5	9012813	3.38	9.92	LOC100784416
	ss715592632	G/A	10	5	9097414	3.38	9.92	glyma05g09390
Palotina	ss715613203	G/A	12	12	5706745	3.93	13.31	Ribonuclease III satt568 satt442 satt192 **
	ss715613207	A/G	12	12	5786241	3.81	12.89	-[d]
Primavera do Leste	ss715613104	A/C	- [c]	12	4670638	3.66	12.33	-
	ss715610817	G/A	13	11	5065170	3.60	10.08	-
	ss715598558	A/G	13	7	6947362	3.49	9.78	Glyma07g076800
	ss715613203	G/A	12	12	5706745	3.21	8.92	-[d]

Chr: Chromosome; LD: Linkage Disequilibrium;
[a] http://soybase.org/snps/
[b] Significant at $-log(P) > 3$
[c] without haplotype
[d] SNPs were associated with the same previous reported QTLs in **

Table S5 - Summary of mixed modeling analyses (Q + K model) for SNPs and haplotypes significantly associated with plant height evaluated in 169 cultivars of soybean in four environment of southern Brazil

Environment	Marker [a]	SNP [b]	Haplotype Block LD	Chr	Position	$log_{10}(P)$	R^2	Nearby Genes
Cascavel	ss715635468	G/A	42	19	45209801[§]	5.75	17.51/9.72[§]	Sd yld 11-6 ** Pl ht 4-2 Pl ht 13-8 Glyma19g196000
	ss715635454	A/G	42	19	45152186[§]	5.66	14.68/ 9.55	-[d]
	ss715635506	C/T	43	19	45441251[§]	5.56	16.86/27.04	-
	ss715635520	A/G	43	19	45525374[§]	5.53	16.76/30.19	-
	ss715635425	A/C	42	19	45000827[§]	5.32	16.08/9.42	Glyma19g37890 Dt1 gene **
	ss715635477	A/G	42	19	45255796[§]	5.02	15.12/31.44	-[d]
	ss715635494	A/G	43	19	45361938[§]	4.84	14.53/28.99	-
	ss715635433	T/C	42	19	45062248[§]	4.45	13.29/27.71	-[d]
	ss715635403	G/A	42	19	44761515[§]	3.94	11.68/ 9.13	-[d]
	ss715601733	C/T	-[c]	8	39969061	3.75	11.08	-
	ss715633774	T/C	20	19	32194361	3.62	10.66	LOC100789162
	ss715609800	A/G	-	11	26755843	3.55	10.46	-
	ss715581751	C/T	-	2	2920341	3.41	10.01	-
Palotina	ss715632400	G/A	71	18	61175038	3.31	9.72	LOC100787543
	ss715634905	G/T	34	19	39723056	3.26	9.55	LOC100786140
	ss715622494	T/C	45	15	48727813	3.22	9.42	LOC100804065
	ss715585767	A/G	32	3	38862467	3.12	9.13	-
	ss715635276	A/C	38	19	43117852	4.36	27.70	LOC100777767
	ss715635224	G/A	-	19	42459502	4.04	27.26	-
	ss715603983	A/G	24	9	38013391	3.15	23.77	-
Primavera do Leste	ss715619979	G/A	21	14	8186078	3.67	12.35	-
	ss715637988	G/A	24	20	37857633	3.58	12.01	LOC100810047
	ss715637964	T/C	23	20	37410040	3.58	12.01	-
	ss715637991	G/A	24	20	37909306	3.42	11.43	-
	ss715619968	T/G	21	14	8128492	3.36	11.22	LOC100804944
Rio Verde	ss715592226	T/C	40	5	41638179	3.81	17.92	-
	ss715592240	C/T	40	5	41740936	3.81	17.92	LOC100788304
	ss715592231	C/T	40	5	41658399	3.27	15.15	-

Chr: Chromosome; LD: Linkage Disequilibrium; [a] http://soybase.org/snps/; [b] Significant at $-log(P) > 3$; [c] without haplotype; [d] SNPs were associated with the same previous reported QTLs in ** [§] SNP associated in Palotina too. R^2 for SNPs associated in Cascavel/Palotina.

Table S6 - Descriptive statistics of phenotypic variation, heritability (h^2) across environments and variance components (G and G × E) of seed yield (SY), seed weight (SW), plant height (PH), days to maturity (DTM) and flowering (DTF) of 141 cultivars of soybean evaluated in eight environments

Trait	Environment	Mean	SD	Min	Max	G	G × E	h^2 (%)
SY (kg ha^{-1})	Env1	2457.59	820.92	806.00	6563.00	75068	351055	56.7
	Env2	1910.82	767.17	233.00	4372.00			
	Env3	1863.71	623.59	125.00	5127.00			
	Env4	670.23	305.47	128.00	1780.00			
	Env5	3319.00	1297.25	176.00	7149.00			
	Env6	1442.93	667.79	299.00	3669.00			
	Env7	1559.18	814.61	136.00	4284.00			
	Env8	1775.69	800.04	152.00	4916.00			
	Mean							
SW (100seed gr)	Env1	12.08	2.31	7.90	25.50	1.50	2.05	81.7
	Env2	12.59	1.99	9.00	25.80			
	Env3	13.49	2.26	7.90	25.00			
	Env4	11.96	1.69	8.20	23.90			
	Env5	12.33	3.15	6.30	19.40			
	Env6	12.30	1.92	7.60	18.40			
	Env7	15.50	1.93	10.30	21.00			
	Env8	14.78	1.96	10.10	21.80			
	Mean							
PH (cm)	Env1	103.45	19.89	55.00	220.00	209.30	101.83	93.4
	Env2	48.36	11.98	20.00	90.00			
	Env3	97.59	19.54	45.00	205.00			
	Env4	38.27	11.59	20.00	75.00			
	Env5	90.34	24.45	30.00	180.00			
	Env6	74.25	23.04	30.00	130.00			
	Env7	46.17	13.72	20.00	95.00			
	Env8	55.52	18.50	23.00	100.00			
	Mean							
DTF (days)	Env1	46.16	10.38	28.00	80.00	44.63	18.42	94.6
	Env2	30.29	5.90	24.00	52.00			
	Env3	47.75	9.41	29.00	82.00			
	Env4	40.49	7.31	28.00	77.00			
	Env5	46.58	10.85	26.00	76.00			
	Env6	46.76	6.89	32.00	70.00			
	Env7	37.39	7.26	24.00	54.00			
	Env8	31.42	6.09	25.00	46.00			
	Mean							
DTM (days)	Env1	126.33	15.39	104.00	256.00	82.00	53.57	91.7
	Env2	88.83	9.84	40.00	172.00			
	Env3	124.89	15.66	97.00	248.00			
	Env4	99.02	13.71	82.00	182.00			
	Env5	133.89	10.59	106.00	164.00			
	Env6	119.59	6.48	106.00	138.00			
	Env7	104.98	9.38	80.00	123.00			
	Env8	98.27	5.56	75.00	123.00			
	Mean							

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

Table S7 - Pearson correlation coefficients among mean variables for seed yield, seed weight, plant height, days to maturity and days to flowering by environment for 141 cultivars of soybean

Environment	Trait	SY	SW	PH	DTM	DTF
	SY	-				
	SW	0.51***	-			
Env1	PH	-0.01ns	0.12*	-		
	DTM	-0.04ns	0.19***	0.63***	-	
	DTF	-0.44***	-0.11ns	0.51***	0.75***	-
	SY	-				
	SW	0.15*	-			
Env2	PH	0.35***	-0.03ns	-		
	DTM	0.39***	0.19**	0.49***	-	
	DTF	0.16**	-0.18**	0.48***	0.65***	-
	SY	-				
	SW	0.28***	-			
Env3	PH	0.09ns	0.26***	-		
	DTM	0.08ns	0.47***	0.61***	-	
	DTF	-0.08ns	0.33***	0.48***	0.79***	-
	SY	-				
	SW	0.26***	-			
Env4	PH	0.23***	0.12ns	-		
	DTM	0.17**	0.26***	0.68***	-	
	DTF	0.04ns	0.15*	0.65***	0.84***	-
	SY	-				
	SW	0.58***	-			
Env5	PH	-0.18**	-0.42***	-		
	DTM	-0.65***	-0.66***	0.35***	-	
	DTF	-0.42***	-0.42***	0.22***	0.44***	-
	SY	-				
	SW	-0.01ns	-			
Env6	PH	0.47***	-0.36***	-		
	DTM	-0.34***	-0.29***	0.08ns	-	
	DTF	-0.08ns	-0.47***	0.37***	0.54***	-
	SY	-				
	SW	0.31***	-			
Env7	PH	0.43***	0.06ns	-		
	DTM	0.39***	0.51***	0.21**	-	
	DTF	0.46***	0.32***	0.13*	0.69***	-
	SY	-				
	SW	-0.01ns	-			
Env8	PH	0.47***	-0.21**	-		
	DTM	0.26***	-0.18*	0.37***	-	
	DTF	0.30***	-0.21**	0.45***	0.66***	-

*Significant at $P < 10^{-2}$
** Significant at $P < 10^{-3}$
*** Significant at $P < 10^{-4}$

Table S8 - Goodness of fit of three different GWAS models for: seed yield (SY), 100-seed weight (SW), plant height (PH), days to flowering (DTF) and days to maturity (DTM) in 141 varieties of soybean evaluated in eight environments of Brazil. Q represents the model with population structure effect; K is the model with kinship effect and Q + K represent the model with the joint effects

Environments	Models	SY		SW		PH		DTF		DTM	
		-Log (L)	BIC	-Log (L)	BIC	-Log (L)	BIC	-Log (L)	BIC	-Log (L)	BIC
Env1	Q	2125.02	2144.82	588.58	608.38	1163.00	1182.79	995.44	1015.23	1095.89	1115.69
	K	2194.71	2209.55	575.02	589.87	1189.97	1204.81	917.82	932.66	1083.07	1097.92
	Q + K	2074.31	2299.06	546.50	571.24	1132.42	1157.16	870.12	894.86	1031.37	1056.12
Env2	Q	2104.15	2123.95	520.84	540.64	872.98	892.27	836.57	856.37	864.56	883.87
	K	2199.40	2214.25	525.58	540.43	920.11	934.57	789.62	804.47	890.59	905.07
	Q + K	2073.94	2098.68	493.25	517.99	861.88	885.98	751.58	776.32	843.52	867.67
Env3	Q	2013.91	2033.71	563.01	582.81	1160.40	1180.19	967.82	987.62	1106.57	1126.36
	K	2128.85	2143.70	580.16	595.01	1187.58	1202.42	883.01	897.86	1105.20	1120.04
	Q + K	2007.98	2032.72	542.90	567.64	1125.28	1150.02	839.03	863.78	1051.55	1076.30
Env4	Q	1792.14	1811.88	479.17	498.91	1006.37	1026.16	884.66	904.45	1063.81	1083.60
	K	1905.23	1920.04	473.52	488.33	1020.99	1035.84	856.71	871.55	1054.77	1069.61
	Q + K	1792.14	1816.81	440.68	465.35	963.88	988.62	814.17	838.91	1000.14	1024.89
Env5	Q	2235.45	2255.19	663.07	682.81	1209.29	1229.05	1009.28	1029.07	954.05	973.67
	K	2308.09	2322.89	652.52	667.32	1264.51	1279.34	1010.36	1025.21	954.63	969.34
	Q + K	2181.67	2206.34	619.95	644.62	1195.69	1220.40	958.47	983.21	902.63	927.16
Env6	Q	1523.98	1542.64	390.88	409.54	857.85	876.39	880.93	900.72	618.95	637.49
	K	1638.61	1652.60	404.83	418.82	890.23	904.14	849.55	864.40	634.32	648.23
	Q + K	1517.16	1540.48	376.58	399.90	821.82	844.99	803.29	828.03	587.98	611.07
Env7	Q	2074.31	2094.07	523.52	543.29	902.09	921.30	892.91	912.71	823.28	842.50
	K	2171.43	2186.26	508.55	523.38	950.20	964.62	840.87	855.72	820.66	835.07
	Q + K	2046.43	2071.14	480.68	505.39	887.56	911.59	798.45	823.19	766.65	790.67
Env8	Q	1948.53	1968.03	486.22	505.73	1044.72	1064.20	821.51	841.30	784.30	803.83
	K	2067.76	2082.38	505.86	520.49	1084.57	1099.17	743.75	758.60	786.05	800.70
	Q + K	1939.14	1963.52	478.26	502.64	1021.78	1046.12	707.93	732.67	736.40	760.81

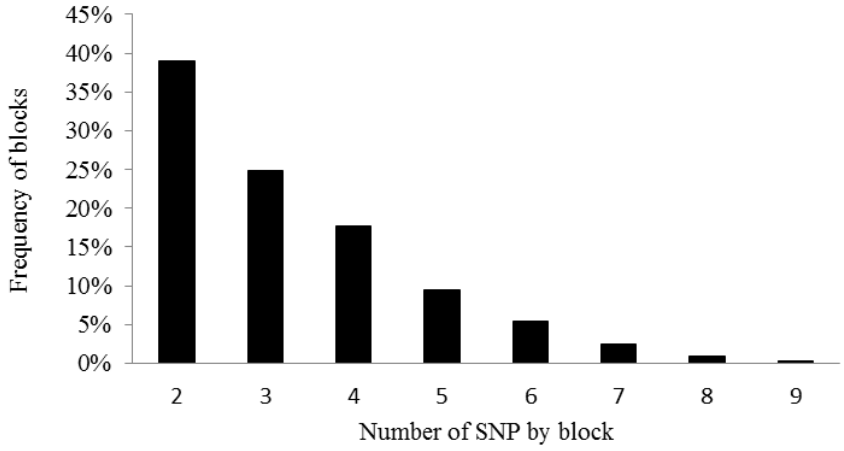


Figure S1 – Frequency of SNPs markers in linkage disequilibrium by block

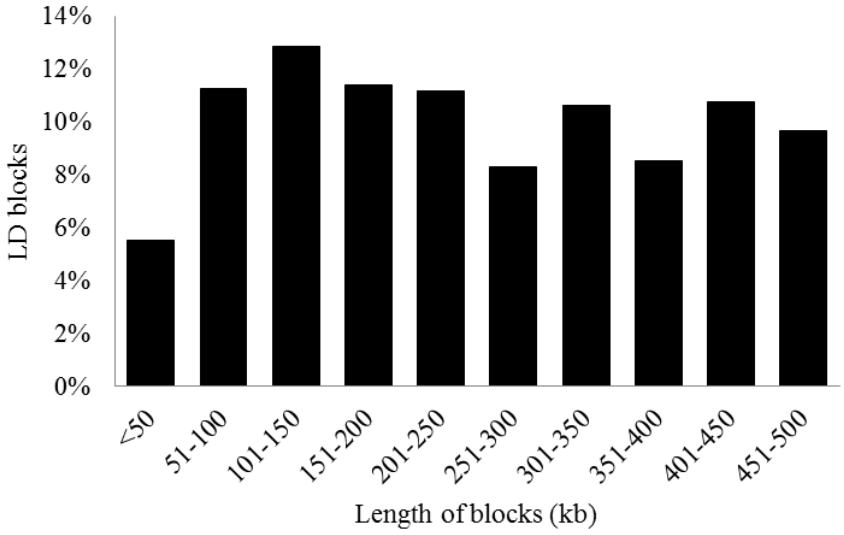


Figure S2 – Frequency of sizes of linkage disequilibrium blocks

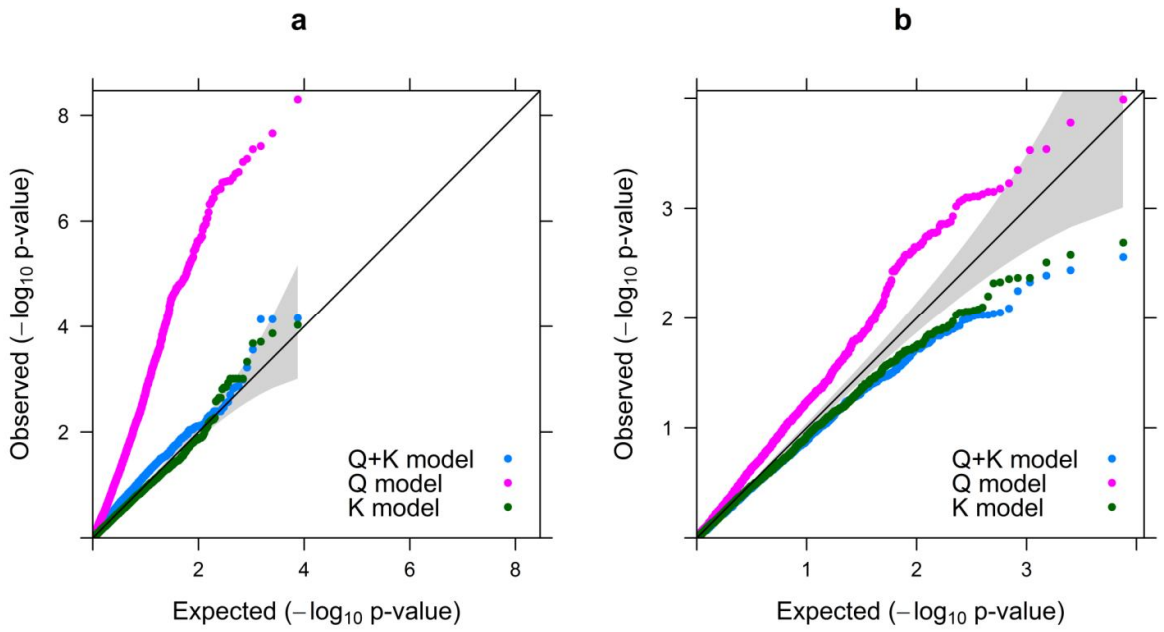


Figure S3 - QQ-plot of MLM comparison for SY in soybean. a) Cumulative distribution of p-values for the Q model, K model and Q + K model for Cascavel environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Palotina environment.

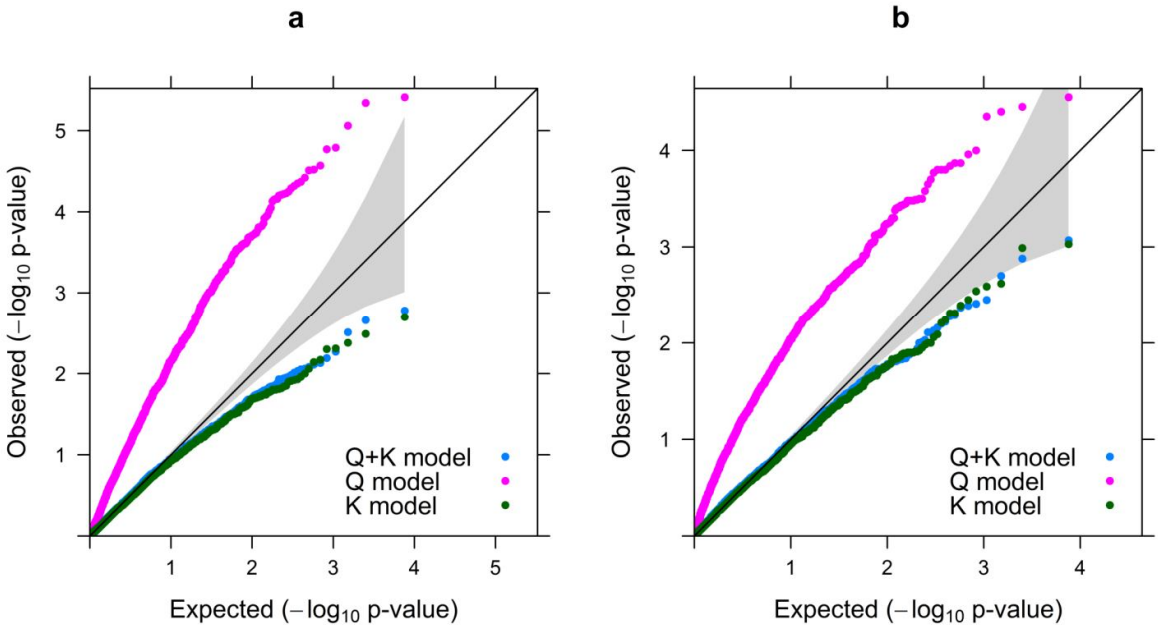


Figure S4 **-** QQ-plot of MLM comparison for SY in soybean. a) Cumulative distribution of p-values for the Q model, K model and Q + K model for Primavera do Leste environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Rio verde environment.

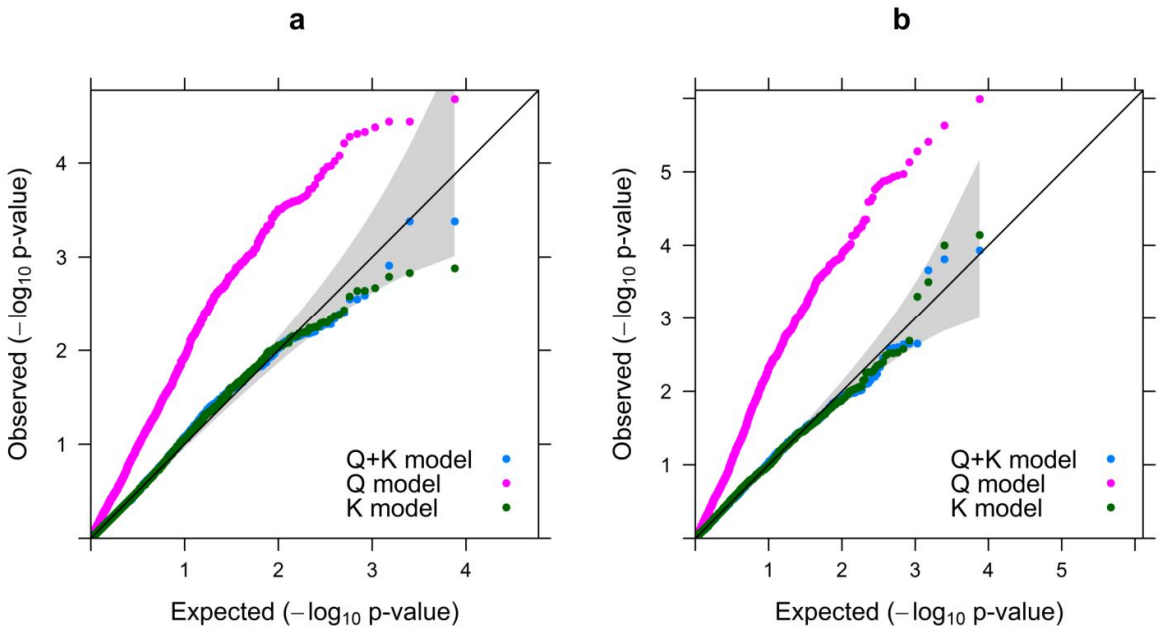


Figure S5 - QQ-plot of MLM comparison for SW in soybean. a) Cumulative distribution of p-values for the Q model, K model and Q + K model for Cascavel environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Palotina environment.

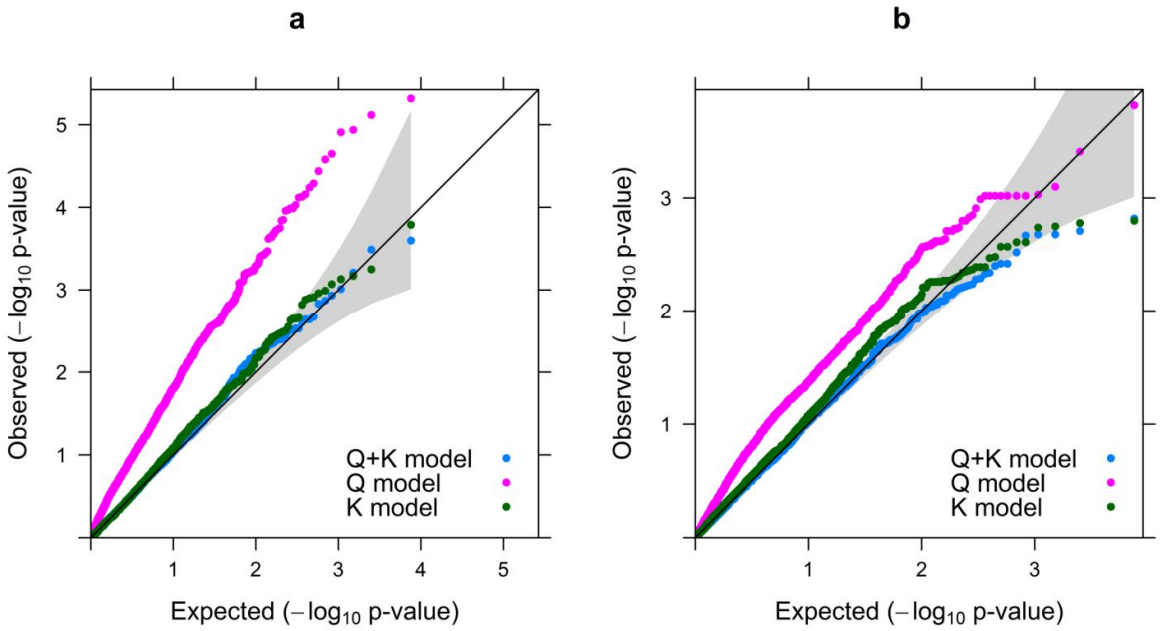


Figure S6 - QQ-plot of MLM comparison for SW in soybean. a) Cumulative distribution of p-values for the Q model, K model and Q + K model for Primavera do Leste environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Rio verde environment.

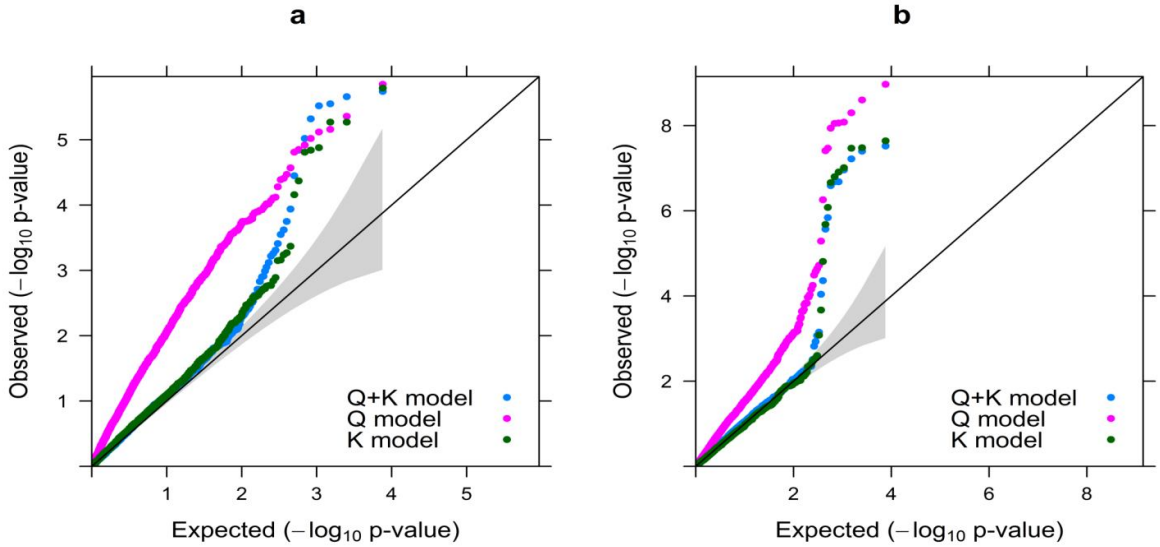


Figure S7 - QQ-plot of MLM comparison for PH in soybean. a) Cumulative distribution of p-values of Q model, K model and Q + K model for Cascavel environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Palotina environment.

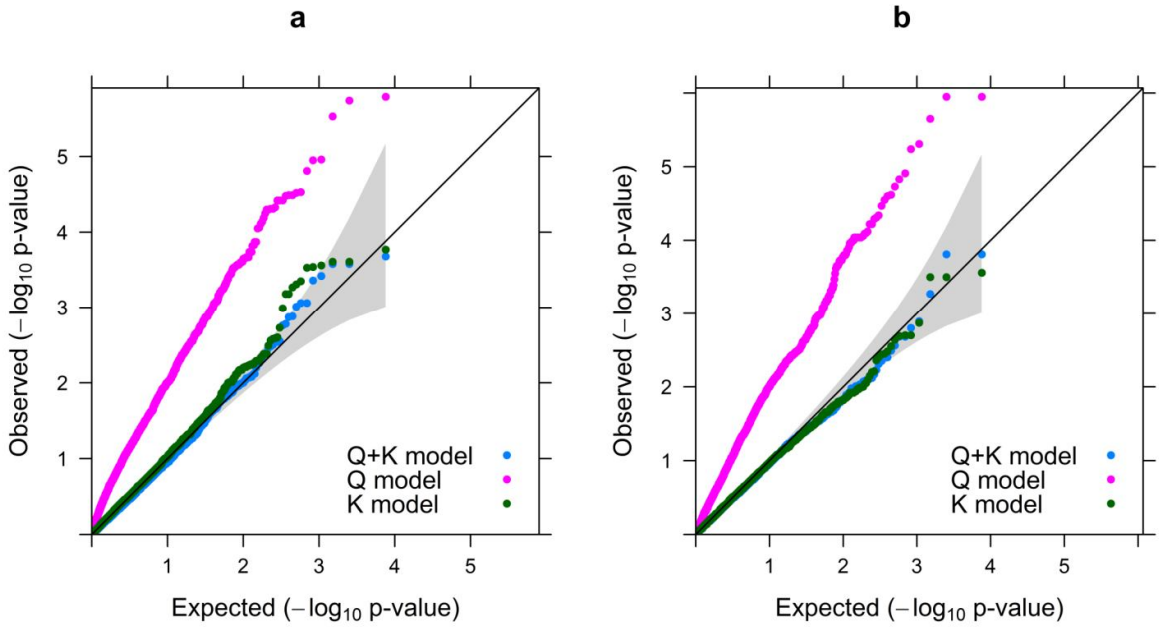


Figure S8 - QQ-plot of MLM comparison for PH in soybean. a) Cumulative distribution of p-values of Q model, K model and Q + K model for Primavera do Leste environment. b) Cumulative distribution of p-values for the Q model, K model and Q + K model for Rio verde environment.

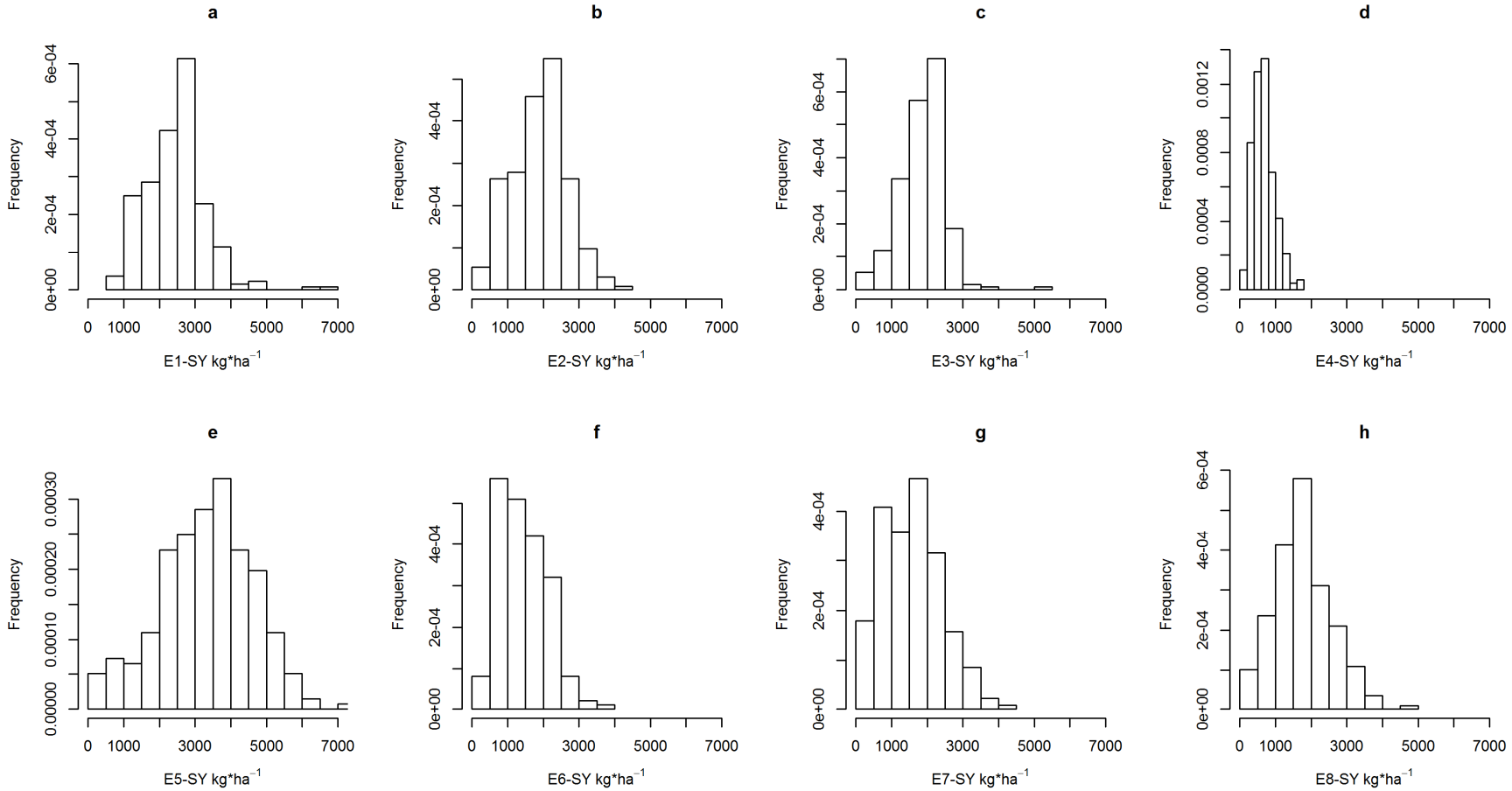


Figure S9 - Relative frequency distribution of observations for seed yield (SY) in 141 cultivars of soybean by environment E1 (a) to E8 (h).

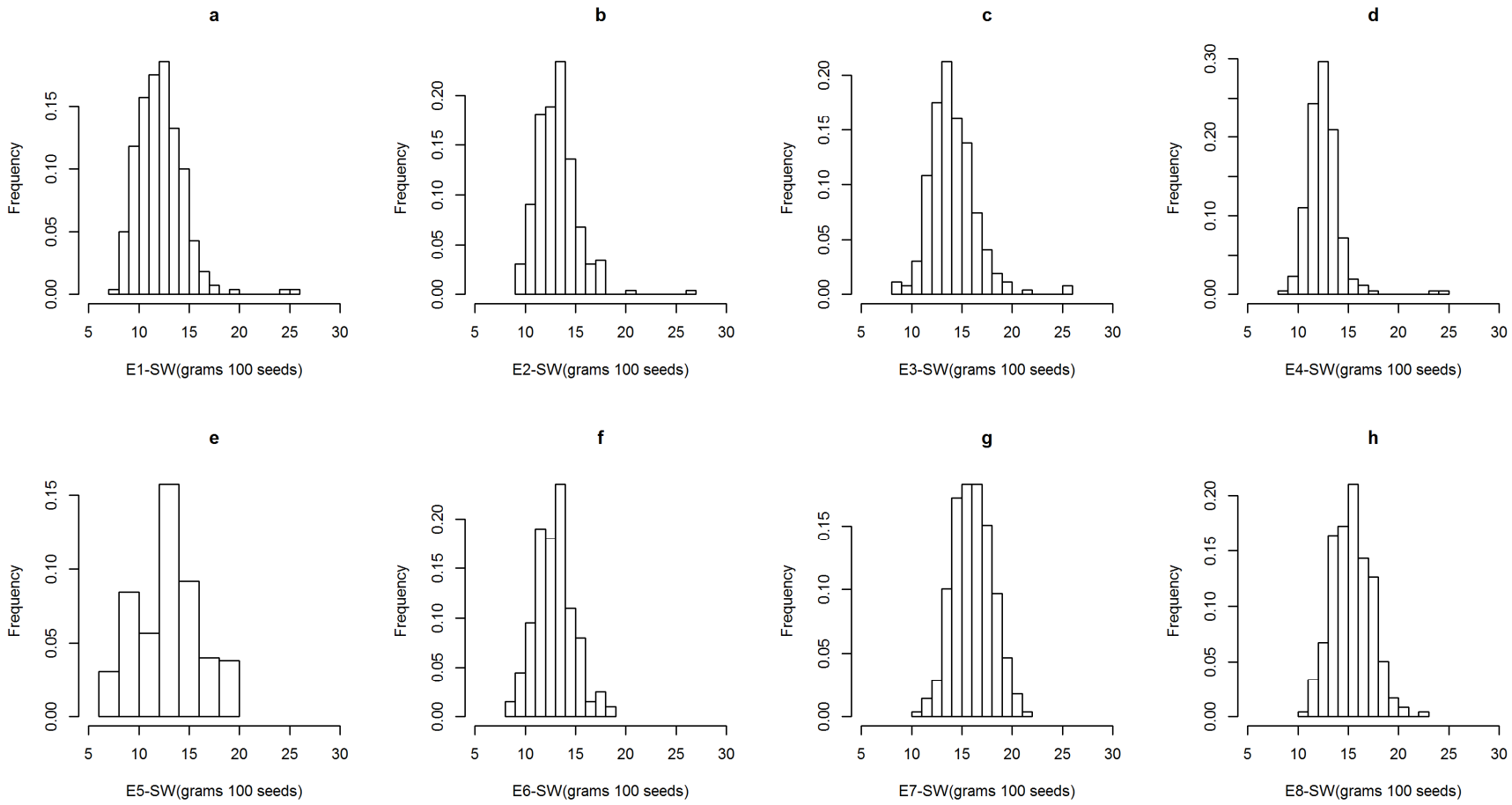


Figure S10 - Relative frequency distribution of observations for 100-seed weight (SW) in 141 cultivars of soybean by environment E1 (a) to E8 (h).

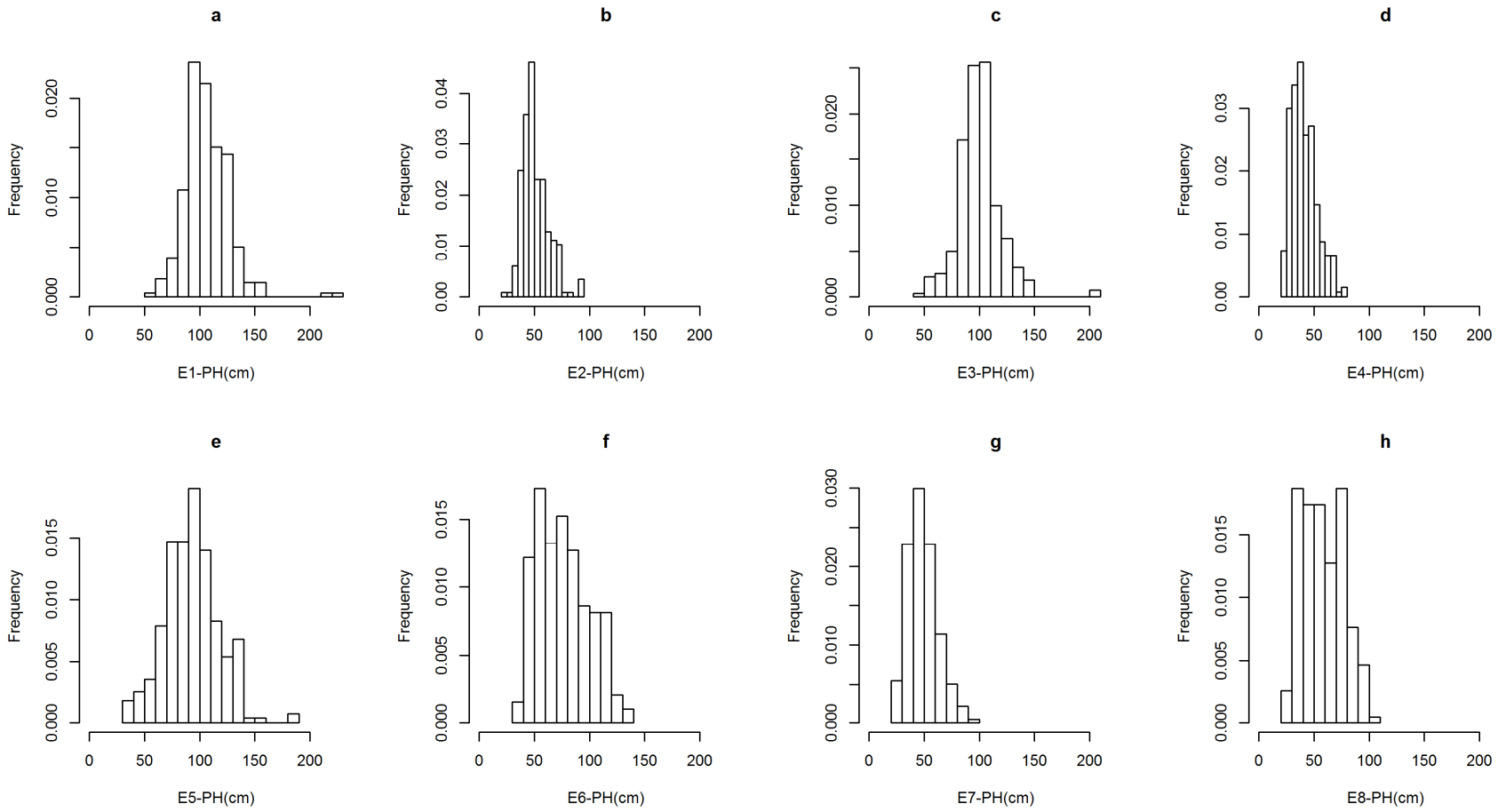


Figure S11 - Relative frequency distribution of observations for plant height (PH) in 141 cultivars of soybean by environment Env1 (a) to Env8 (h).

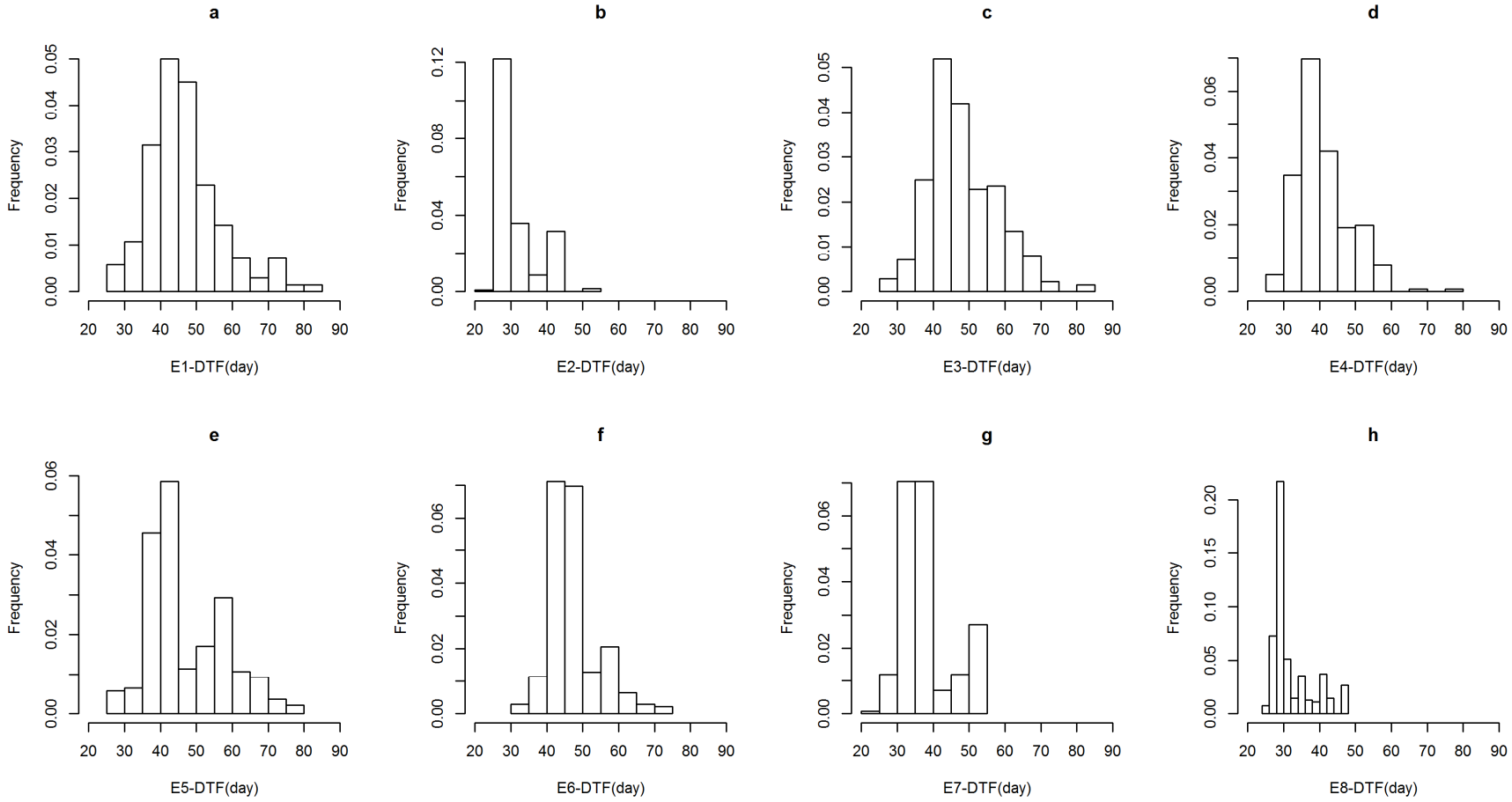


Figure S12 - Relative frequency distribution of observations for days to flowering (DTF) in 141 cultivars of soybean by environment Env1 (a) to Env8 (h).

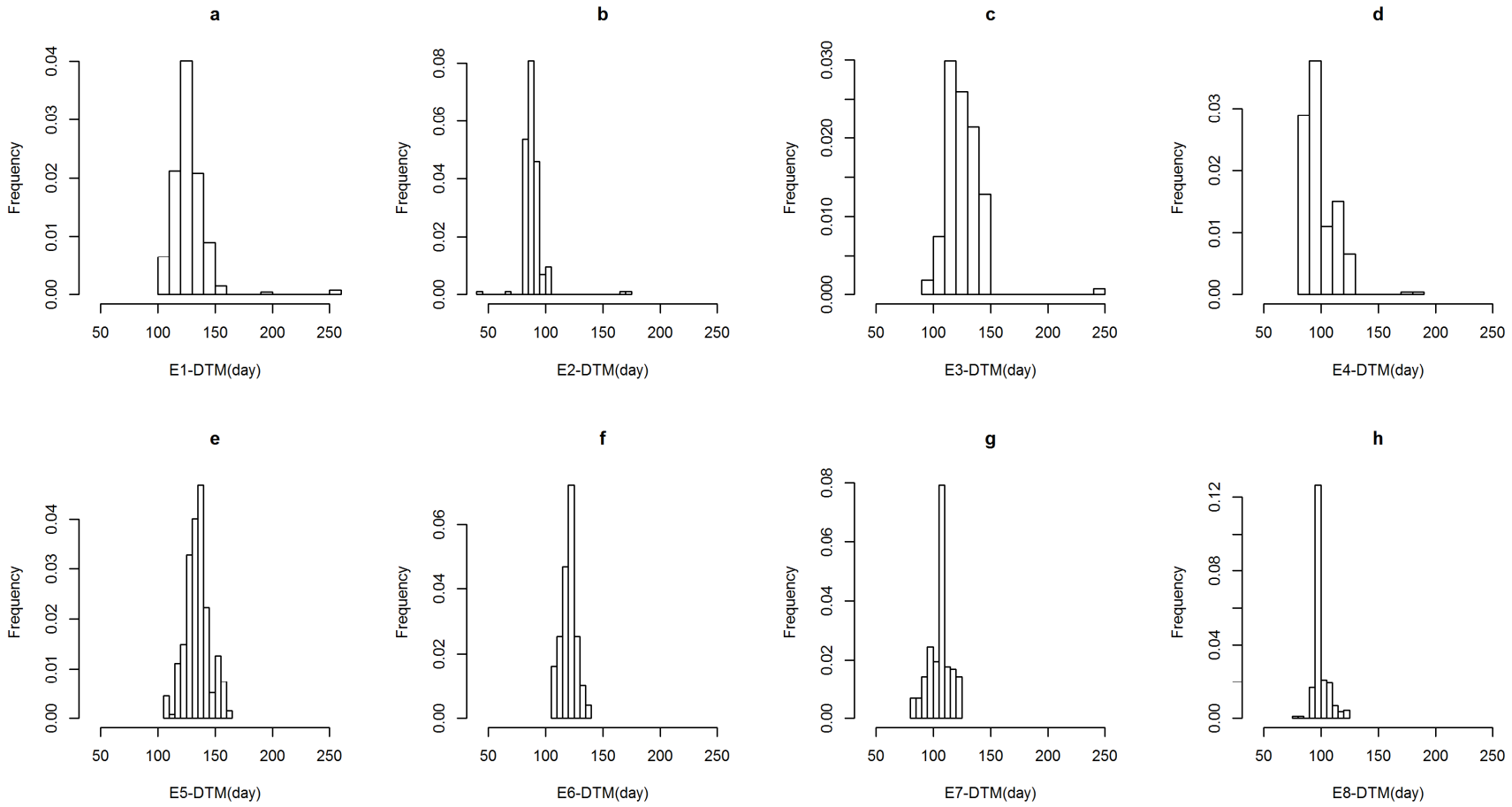


Figure S13 - Relative frequency distribution of observations for days to maturity (DTM) in 141 cultivars of soybean by environment Env1 (a) to Env8 (h).