Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Mapeamento associativo e estrutura populacional em germoplasma exótico de soja

Mônica Christina Ferreira

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

Piracicaba
2015

Mônica Christina Ferreira
Engenheira Agrônoma

Mapeamento associativo e estrutura populacional em germoplasma exótico de soja

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2015

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

Ferreira, Mônica Christina
Mapeamento associativo e estrutura populacional em germoplasma exótico de soja / Mônica Christina Ferreira. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2015.
107 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura “Luiz de Queiroz”.

1. *Glycine max* 2. Produtividade de grãos 3. Mapeamento associativo 4. Diversidade genética I. Título

CDD 633.34
F383m

AGRADECIMENTOS

À Deus responsável por todas as graças e bênçãos em minha vida.

Aos meus pais, Marcos e Mirna, por sempre me apoiarem em todas às escolhas na minha vida, pelo esforços e sacrifícios para que eu conseguisse completar toda a minha formação acadêmica.

À toda minha família que sempre me motivou, principalmente aos meus avós, Rosa, Belchior, Maria José e Benedito, que tanto ajudaram nos momentos mais difíceis.

Ao meu professor e orientador José Baldin Pinheiro pela confiança creditada a mim para condução deste trabalho, pelos conselhos e ajuda e principalmente por querer me orientar, sou imensamente grata.

Ao meu amigo e namorado Diego Lopes, por todo sacrifício e incentivo nestes quatro anos para que pudesse concluir minha formação em outra cidade e até mesmo em outro país.

Aos colegas do grupo LAB-DGM, Kênia Oliveira, Fabiani da Rocha, Fabiana Freitas Moreira, Felipe Bermudez, Diane Simon Rozzetto, Ellida Silvestre Aguiar, Camila Câmpelo, João Paulo Gomes Viana, Maria Imaculada Zucchi, Vanessa Rizzi, Miklos Maximiliano Bajay, Marcos Siqueira, Carlos Eduardo, Fátima Bosetti, Jaqueline Campos, Eleonora Zambrano, Alessandro Alves, Patrícia Sanae, Maisa Curtolo, Mariana Novello, Carolina Grando, Matheus Dominiquini, Jéssica Gimenez Tâmbalo, Júlia Morosini, Nancy Farfan, Sabrina Della Bruna, Maurício Terasawa e Milene Moller, pelo companheirismo, momentos de descontração, ajudas e ensinamentos, me sinto parte desta família. Em especial ao grupo SOJA, pela imensa ajuda na condução dos experimentos.

Aos colegas da pós graduação em Genética e Melhoramento de Plantas. Em especial a Melina Teixeira, Marcela Mendes, Fernanda Aparecida, Karina Lima, Evellyn Couto, Glaucia, Hendrie, Iradenia, Paolo, Thiago e Nelson, pela parceria nas aulas e pelos momentos especiais de descontração.

A todos os funcionários do departamento de genética da ESALQ, em especial ao Claúdio Segateli por toda ajuda no manejo dos experimentos, além da amizade e conselhos. Aos funcionários de campo Márcio e Domingos Amaral.

Ao Professor Randal Nelson e toda sua equipe pelo suporte durante o meu período de doutorado sanduíche em Illinois.

Aos professores do Departamento de Genética pelos ensinamentos e suporte.

Enfim a todas as pessoas especiais em minha vida, que fizeram e fazem parte desta jornada, eu agradeço profundamente.

Ao professor Magno Antônio Patto Ramalho pelos ensinamentos e principalmente por me introduzir ao melhoramento de plantas, a ciência e por vezes arte a qual sou hoje apaixonada.

SUMÁRIO

RESUMO

Mapeamento associativo e estrutura populacional em germoplasma exótico de soja

A soja é uma das culturas mais importantes do mundo, além de ser a principal *comodity* brasileira. Entretanto apesar dos ganhos crescentes de produtividade a base genética da cultura no país é estreita. Sendo assim, é importante a identificação e caracterização de fontes de variabilidade para os programas de melhoramento de soja. Dado o exposto, os objetivos deste estudo foram *i)* avaliação da produtividade de grãos e caracteres agronômicos correlacionados; *ii)* caracterização da diversidade fenotípica; *iii)* caracterização da diversidade genética e estrutura de populações; e iv) mapeamento associativo para produtividade de grãos. Os acessos foram fenotipados nos anos agrícolas de 2012/2013 e 2013/2014, em cinco ambientes. As características avaliadas foram: altura da planta na maturidade, período de granação, valor agronômico, acamamento, massa de cem sementes, número de dias para a maturidade, inserção da primeira vagem, altura da planta no florescimento, teor de óleo e produtividade de grãos. A análise dos dados fenotípicos foi feita pelo software SELEGEN utilizando modelos mistos, e a árvore de regressão para identificação dos caracteres correlacionados foi feita pelo software JMP SAS. A diversidade fenotípica foi feita a partir de todas as características avaliadas anteriormente utilizando três tipos de análises: os métodos de agrupamento de *Ward* e *Average Linkage* no software Power Marker, e pela análise de componentes principais no software JMP SAS. A genotipagem dos acessos foi realizada pelo Axiom® Soybean Genotyping Array contendo 10017 SNPs polimórficos para os acessos genotipados. A partir dos dados de marcadores foi feita a caracterização da diversidade genética pelo software Power Marker. Além disso, foram realizadas análises de estrutura de população pelo software STRUCTURE e pelo pacote do R, *adegenet*. A análise de associação foi efetuada pelo software TASSEL, utilizando o modelo misto MLM (*Q+K*). Duas abordagens foram utilizadas na análise de associação, a primeira utilizando as médias fenotípicas ajustadas para BLUP dos cinco ambientes e a segunda utilizando apenas as médias de cada local individualmente. Na análise fenotípica os acessos Dowling, PI 417563, PI200526, PI 377573 e PI 159922, apresentaram boa produtividade de grãos nos cincos ambientes avaliados. A caracterização molecular e fenotípica da diversidade indicou a presença de variabilidade genética no painel de acessos avaliados. Além disso, foi possível a identificação de dois grupos (k=2) em ambas as análises de estrutura da população utilizadas. No mapeamento associativo, foram detectadas sete associações marcador-característica com $p<0,001$ e com correção para múltiplos testes $q<0,1$. Dentre estas, quatro foram significativas no modelo de análise conjunta dos cinco ambientes e para o ambiente dois. As demais associações foram significativas somente para este último local.

Palavras-chave: *Glycine max*; Diversidade genética; Desequilíbrio de Ligação; Modelos Mistos

ABSTRACT

Associative mapping and population structure in exotic soybean germplasm

Soybean is one of the most important crops in the world, and is Brazil's main commodity. However despite growing yield gains the genetic basis of culture in the country is narrow. Therefore, the identification and characterization of sources of variability for soybean breeding programs is important. On this basis, the objectives of this study were i) assessment of grain yield and agronomic traits correlated; ii) characterization of the phenotypic diversity; iii) characterization of genetic diversity and population structure; and iv) associative mapping for grain yield. The inbred lines were phenotyped in the agricultural years of 2012/2013 and 2013/2014, in five environments. The traits evaluated were: plant height at maturity, fruit filling period, agronomic value, lodging, mass of hundred seeds, number of days to maturity, first pod, plant height at flowering, oil content and grain yield. The analysis of phenotypic data was made by SELEGEN software using mixed models, and regression tree for identification of correlated traits was made by JMP SAS software. The phenotypic diversity was made from all the features previously evaluated using three types of analysis: the Ward clustering methods and Average Linkage from the Power Marker software, and the principal component analysis in SAS JMP software. Genotyping was performed by Axiom® Soybean Genotyping Array containing 10017 polymorphic SNPs genotyped for the soybean lines. From the markers data was taken the genetic diversity analysis by Power Marker software. In addition, population structure analysis was performed by Structure software and the R package, adegenet. The association analysis was performed by TASSEL software using the mixed model MLM (Q + K). Two approaches were used in the association analysis, the first using the phenotypic average adjusted to BLUP values for the five enviroments and the second one using only the means of each site individually. In the phenotypic analysis the lines: Dowling, PI 417563, PI200526, PI 377573 and PI 159922 showed good grain yield in the five evaluated environments. The molecular and phenotypic characterization of diversity indicated the presence of geneticvariability in the inbreed lines. Moreover, it was possible to identify two groups (k = 2) in both population structure analysis used. In the associative mapping, were detected seven marker-trait associations with $p < 0.001$ and with correction for multiple tests $q < 0.1$. Among these, four were significant in the pooled analysis model with five environments and at the individually environment two. The other variables were significant only for the latter location.

Keywords: *Glycine max*; Genetic diversity; Linkage disequilibrium; Mixed models

1 INTRODUÇÃO

1.1 Importância econômica da soja

O agronegócio brasileiro é responsável por 24% do Produto Interno Bruto (PIB) do País segundo dados do Ministério da Agricultura Pecuária e Abastecimento (MAPA, 2015). O setor foi responsável pelo crescimento da economia em relação ao mesmo período de 2014 com um incremento de 4% em relação ao mesmo período do ano anterior (IBGE, 2015). O principal destaque deste cenário foi a produção de soja (*Glycine max*) com uma área plantada na safra de 2014/2015 de aproximadamente 32 milhões de hectares e produção de 96 milhões de toneladas, dez milhões a mais que a safra anterior (CONAB, 2015).

Além de sua importância na economia nacional, sendo a principal commodity brasileira (CONAB, 2015), a soja é a oleaginosa mais produzida e consumida mundialmente, sua relevância e destaque se deve a ampla gama de produtos obtidos através dos grãos. Seu consumo varia desde alimentação animal na forma de farelo até o consumo humano através de produtos como óleo de cozinha, carne de soja, leite, consumo in natura e outros derivados. Além disto, deve-se destacar seu papel ambiental através de sua utilização para produção de biocombustíveis (EMBRAPA, 2004).

No Brasil a cultura passou a se destacar a partir de década de 70, onde a produção de soja passou a ter grande relevância na agricultura, principalmente devido ao aumento das áreas cultivadas e uso de novas tecnologias que possibilitaram o aumento na produtividade, das quais pode-se destacar as técnicas de manejo do solo e as contribuições do melhoramento genético (BRANDÃO, REZENDE E MARQUES, 2005). Com isto, em 2003, o Brasil se tornou o segundo maior produtor mundial de soja (EMBRAPA, 2004).

Entretanto, apesar das potencialidades do Brasil para produção desta oleaginosa, o país possui desafios a serem enfrentados, como aumento da produção sem que seja necessária a incorporação de novas áreas de plantio, manejo adequado de insetos e doenças sem a utilização demasiada de defensivos agrícolas, além de melhoria das condições de logística, infraestrutura no armazenamento e transporte dos grãos. Desafios estes que, se ultrapassados,

resultariam em uma maior competitividade do complexo de soja brasileiro no agronegócio mundial.

1.2 Recursos Genéticos Vegetais

Os Recursos Genéticos compreendem a variabilidade entre plantas, animais e microrganismos que compõem a biodiversidade e que possuem potencial para utilização comercial em áreas como biotecnologia e melhoramento entre outras áreas afins. Por sua vez, os chamados Recursos Genéticos Vegetais (RGV), compreende apenas aqueles organismos da flora, ou seja, as plantas (NASS et al., 2001). Os RGV apresentam valor imenso para as gerações atuais e futuras de pesquisadores tanto para aqueles envolvidos com melhoramento genético convencional quanto para biotecnologia (ESQUINAS-ALCÁZAR, 1993). Estes recursos podem ser considerados como bancos genéticos nos quais podem ser encontradas soluções para diversos problemas como as mudanças climáticas, resistências ao calor, seca e déficit hídrico, além de estresses bióticos como insetos e doenças (NASS et al., 2001).

Em relação aos RGVs estima-se que tenham sido descritas cerca de 300 mil espécies de plantas e que o homem tenha utilizado aproximadamente 3000 para sua alimentação. Atualmente, sabe-se que são cultivadas ao redor de 300, sendo que destas, apenas 15 respondem por 90% de toda alimentação da população mundial (PATERNIANI, 1988; GOODMAN, 1990). Isto reflete inegavelmente a drástica redução da diversidade genética das espécies e plantas que o homem utiliza, representando a erosão genética que vem ocorrendo ao longo do tempo devido ao processo de seleção e melhoramento (NASS et al., 2001).

Esta baixa utilização da diversidade existente e o estreitamento da base genética das plantas que o homem utiliza para cultivo e alimentação foi responsável por várias situações desastrosas de vulnerabilidade genética na história mundial. Um dos exemplos mais antigos e mais famosos é a “Grande Fome”, que ocorreu na Irlanda entre 1845 e 1851, onde as plantações de batata alimento principal na dieta do país, foram devastadas pelo fungo *Phytophthora infestans*. Aproximadamente 25% da população (cerca de 2 milhões de pessoas) morreram de fome e cerca de 1,5 milhão de irlandeses deixaram o país (WOODHAM-SMITH, 1962). Há também, um exemplo bem mais recente, na década de 70 nos Estados Unidos e Rússia o

fungo *Helminthosporium maydis* devastou as lavouras de milho devido a uniformidade genética dos cultivos (NASS et al., 2001).

Entretanto, apesar dos incidentes relacionados ao estreitamento da base genética das plantas cultivadas, a maioria dos cruzamentos feitos pelos melhoristas continua envolvendo principalmente linhagens elite (GOODMAN, 1990), já que estes materiais apresentam elevado potencial agronômico e econômico além de já serem adaptados ao ambiente em questão (DUVICK, 1984; PATTERNIANI, 1987; TROYER, 1990). Os genes nas coleções de germoplasma são pouco utilizados nos programas de melhoramento, ao menos até que estes sejam previamente avaliados, caracterizados e sejam assim, incorporados aos materiais elites e aos programas de melhoramento de plantas.

1.3 Germoplasma e Base Genética da Soja

A soja cultivada *Glycine max* provavelmente originou e foi domesticada provavelmente da soja selvagem *G. soja*, a qual é nativa de regiões da China, Taiwan, Japão, Coréia e Rússia (HERMANN, 1962; HYMOWITZ, 1970, SINGH & HYMOWITZ, 1999). Contudo não há um estudo sobre em quais dessas regiões a domesticação ocorreu. A referência mais antiga a esta espécie foi feita pelo imperador Chinês Shen Nung em um herbário em 2838 a. C. Acredita-se que com o crescimento do comércio, a soja tenha sido levada para os demais países (BONETTI, 1981).

Atualmente a soja é cultivada em diversas regiões do globo, entretanto, somente uma pequena fração da diversidade genética disponível está sendo correntemente utilizada, no melhoramento da soja (CARTER et. al., 2004). Os melhoristas estão restritos em explorar somente os recursos de apenas um grupo de maturação e esta é a razão do estreitamento da base genética desta espécie (GIZLICE et al., 1996; BURTON, 1997; SINGH & HYMOWITZ, 1999; SINGH et al., 2007b). Hartwig (1973) e St. Martin (1982) estimaram que os melhoristas obtiveram progresso em seus programas ao custo da perda substancial da variabilidade genética disponível.

Segundo Delannay et al. (1983), o germoplasma de soja do Brasil e EUA possuem origens próximas, sendo o brasileiro derivado na maior parte do americano. Hiromoto & Vello (1986) caracterizam pioneiramente a base genética da

soja brasileira, segundo os autores, 11 ancestrais explicaram 89% da base genética da cultura. Estes dados já indicavam naquela época a base genética estreita desta leguminosa. Em trabalho realizado por Priolli et al. (2004), avaliando a diversidade da soja cultivada no Brasil nas três últimas décadas, utilizando dados oriundos de marcadores microsatélites, obtiveram resultados que indicaram que o germoplasma utilizado nos programas de melhoramento brasileiro mantiveram um nível constante de diversidade genética nos últimos trinta anos. Por sua vez, Wysmierski (2010) utilizou dados de genealogia para avaliar a contribuição de ancestrais à cultura da soja e obteve resultados diferentes desta última, indicando que a base da soja é bastante estreita e que, apesar da incorporação de novos genótipos, houve um estreitamento da base genética da cultura ao longo dos anos.

Os bancos de germoplasma são fonte de diversidade genética para os melhoristas de plantas, ou seja, os recursos genéticos estão disponíveis para melhorar uma espécie cultivada ou com potencial uso para agricultura. De acordo com dados coletados pelo IIRG (Instituto Internacional de Recursos Genéticos) em 2001, mais de 170 mil acessos de *G. max* são mantidos por mais de 160 instituições ao redor do mundo. A China possui a maior coleção de germoplasma de soja coletado, com quase vinte e seis mil acessos (WANG, 1982; CHANG e SUN, 1991). O segundo maior banco de germoplasma é a coleção do USDA (*United States Department of Agriculture*) com um pouco mais de dezoito mil acessos. O Brasil conta com duas coleções de germoplasma de soja, uma no CERNARGEN (Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia) com quase quatro mil e setecentos acessos, e a outra na EMBRAPA Soja (CARTER et. al., 2004).

A partir de acessos selecionados do banco de germoplasma da EMBRAPA SOJA, Mulato et al. (2010) avaliou a diversidade genética de setenta e nove acessos de vários locais. Para o desenvolvimento do estudo foram utilizados marcadores agro morfológicos e microssatélites. Os resultados indicaram uma quantidade bastante significativa de alelos raros e alguns genótipos contando com alelos exclusivos, a diversidade genética encontrada foi alta e exibiu um nível moderado de associação entre a divergência genética e a origem geográfica.

1.4 Mapeamento Associativo

O mapeamento associativo também conhecido como mapeamento por desequilíbrio de ligação ou fase gamética visa descobrir associações entre marcadores moleculares e características fenotípicas. Este tipo de mapeamento se baseia no desequilíbrio de ligação (DL) de uma população o qual pode ser definido como a associação não aleatória de alelos de diferentes locos (FLINTGARCIA et al., 2003). Assim, este tipo de mapeamento tenta utilizar a variação contida em uma população para identificar associações entre um gene relacionado à característica de importância a um marcador molecular (WANG et al., 2008).

A análise de associação possui muito em comum com o mapeamento de QTLs. Ambos tentam identificar, via inferência estatística, a co-segregação de marcadores genéticos polimórficos com os genes envolvidos na variação da característica. Entretanto, os dois métodos diferem em algumas propriedades chave, as quais têm implicações nas aplicações de cada uma dessas técnicas. O mapeamento de QTL geralmente envolve populações estruturadas, espécies de plantas com geração curta, mapeamento com populações advindas de linhagens homozigotas são comumente utilizados, além de espécies advindas de cruzamentos com parentesco conhecido. O resultado do uso populações que possuem poucas gerações de recombinação é a maximização do DL por par de bases. Portanto, marcadores relativamente distantes podem co-segregar com os QTLs. Em contraste, populações não estruturadas utilizadas nos estudos de associação, possuem muitas gerações de descendentes de um ancestral comum, portanto, foram sujeitos a muitos eventos de recombinação e devido a isto, somente os DL fortemente ligados serão detectados, o que implica em alta associação entre um marcador e um QTL (ORAGUZIE & WILCOX, 2007).

Há um grande número de índices para mensurar o desequilíbrio de ligação, sendo os mais utilizados as estatísticas r^2 e D’ (FLINT-GARCIA et al., 2003). Entretanto, a estatística D’ é muito influenciada por tamanhos pequenos da população, além disto, a estatística r^2 é mais conveniente por ser um indicativo de como os marcadores podem ser associados com a região do genoma responsável pela característica de interesse (FLINT-GARCIA et al., 2003).

A resolução típica observada em estudos genéticos utilizando linhas puras recombinantes é de 10-30 cM (ALPERT e TANKSLEY, 1996; STUBER et al., 1999).

Com esta resolução (equivalente a 10-30 milhões de pares de bases) centenas de genes dentro do QTL ainda permanecerão sem identificação. Estudos de associação baseados no DL permitem a identificação de quais genes são representados por estes QTLs. Somente os polimorfismos extremamente próximos aos locos com efeitos fenotípicos são prováveis de serem significativamente associados com um caráter em uma população tipicamente utilizada no mapeamento associativo, fornecendo uma resolução muito mais fina do que os demais mapeamentos de QTL (REMINGTON et al., 2001).

Há duas formas de se realizar o mapeamento associativo, o mapeamento em todo o genoma ou mapeamento a partir de genes candidatos. No primeiro caso é utilizado um grande número de marcadores juntamente com dados de fenotipagem da característica de interesse, e todos os genes são avaliados simultaneamente. Na segunda abordagem, de genes candidatos, a genotipagem é feita em regiões especificas do genoma que contenham os genes de interesse (GEBHARDT, 2007). O tipo de abordagem a ser utilizado depende do nível do desequilíbrio de ligação da população a ser estudada, se o DL decair a uma distância curta no cromossomo há necessidade do uso de um maior número de marcadores para detectar as associações, se o contrário for verdadeiro será necessária uma menor quantidade de marcadores (FLINT-GARCIA et al., 2003; MORGANTE e SALAMINI, 2003; MALOSETTI et al., 2007).

O principal problema no mapeamento associativo está em encontrar associações falso positivas devido a estrutura da população estudada (PRITCHARD, 2001). Isto ocorre devido a sub-estruturação ou subdivisão dentro de uma população e como consequência disto podem ser encontrados correlações entre locos não associados (MACKAY e POWELL, 2007). Para corrigir este problema há métodos estatísticos que possibilitam a identificação de grupos dentro de uma população (PRITCHARD et al., 2000). Além disto, é indicado o uso de dados fenotípicos acumulados ao longo dos anos, ou seja, vários ambientes, e neste caso os modelos mistos podem ser utilizados como um poderoso aliado ao mapeamento associativo. Este método tem sido relatado como muito eficiente na diminuição de falsas associações entre marcadores e caracteres fenotípicos (MALOSETTI et al., 2007).

O mapeamento por desequilíbrio de ligação pode ser utilizado em diversos tipos de populações como coleções de germoplasma, genótipos elite de um

programa de melhoramento genético, populações naturais dentre outras (GEBHARDT et al., 2004).

Em soja os estudos de associação já tiveram êxito no estudo de caracteres complexos, como clorose devido à deficiência de ferro (WANG et al., 2008), parâmetros relacionados a clorofila (HAO et al., 2012) e, teor de proteína na semente (JUN et al., 2007). Hao et al (2012) identificou 19 SNPs (polimorfismo de uma única base) e 5 haplótipos associados com a produtividade em soja utilizando análise de associação em todo o genoma e os marcadores foram localizados dentro ou próximos a QTLs previamente reportados. Wen et al., (2014) trabalhando com dois painéis de soja elite dos EUA e 5361 SNPs identificou 20 locos associados a doença da morte súbita da soja (SDS), dentre estes locos encontrados, sete já foram descritos previamente na literatura. Também trabalhando com resistência a doenças Iquira et al., (2015) trabalhando com genotipagem por sequenciamento (GBS) e um painel de mapeamento composto por 101 PI's (*Plant Introductions*), os autores encontraram três regiões do genoma associadas a resistência ao Mofo Branco.

Para a produtividade de grãos trabalhos envolvendo mapeamento associativo também podem ser encontrados na literatura. Hao et al. (2011), utilizou o mapeamento associativo com marcadores SNPs em 191 acessos de soja selvagem e 5 ambientes diferentes, para detectar associações para produtividade de grãos e características correlacionadas. Os autores descobriram 19 SNPs e 5 haplotipos associados a produtividade de grãos e seus componentes. Hu et al. (2014) trabalhando com 113 linhagens de sojas selvagens e marcadores microssatélites, encontrou 5 associações pra produtividade de grãos e outras características.

Referências

ALPERT, K. B.; GRANDILLO, S.; TANKSLEY, S. D. High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw 2.2: a major QTL controlling fruit weight is common to both red- and green-fruited tomato species. **Proceedings Of The National Academy Of Sciences Of The United States Of America**, Washington, v. 91-91, n. 6-7, p.15503-15507, 1995.

BONETTI, L. P. Distribuição da soja no mundo : origem, história e distribuição. In : MIYASAKA, S.; MEDINA, J.C. (Ed.). **A soja no Brasil**. Campinas : ITAL, p. 1-6, 1981.

BRANDÃO, A. S. P.; REZENDE, G. C.; MARQUES, R. W. C. **Crescimento agrícola no período 1999/2004, explosão da área plantada com soja e meio ambiente no Brasil**. Rio de Janeiro: IPEA, Texto para Discussão nº 1062, 2005.

BURTON, Joseph W. Soyabean (Glycine max (L.) Merr.). **Field Crops Research**, Amsterdam, v. 53, n. 1-3, p.171-186, 1997.

CARTER, T. E.; NELSON, R. L.; SNELLER, C. H.; CUI, Z. Genetic Diversity in Soybean. In: BOERMA, H. R. (Ed.); SPECHT, J. E. Soybean: **improvement, production and uses**. 3. ed. Madison: American Society Of Agronomy, 2004. Cap. 8. p. 303-396.

CHANG R. Z.; SUN, J. Y. Catalogues of Chinese Soybean Germplasm and Resources: Continuation I. **China Agricultural Press**, Beijing, 1991.

DELANNAY, X.; RODGERS, D. M.; PALMER, R. G. Relative Genetic Contributions Among Ancestral Lines to North American Soybean Cultivars. **Crop Science Society Of America**, Madison, v. 23, n. 5, p.944-949, 1983.

DUVICK, D.n.. Genetic contribution to yield gains of U.S. hybrid maize, 1930 to 1980. In: FEHR, W.r. (Ed.). **Genetic contributions to yield gains of five major crop plants**. Madison, Crop Science Society Of America, 1984, p. 15-47.

EMBRAPA . **Tecnologias de Produção de Soja Região Central do Brasil 2004**: A Soja no Brasil. Paraná, 2004. Disponível em: <http://www.cnpso.embrapa.br/producaosoja/SojanoBrasil.htm>. Acesso em: 20 ago. 2015.

ESQUINAS-ALCAZAR, J.T. Plant genetic resources. In: HAYWARD, M.d.; BOSEMARK, N.O.; ROMAGOSA, I. **Plant Breeding: Principles and Prospects**. Londres: Chapman & Hall, 1993. p. 33-51.

FLINTGARCIA, S. A.; S. A.; THORNSBERRY, J. M.; BUCKLER, E. S. Structure of linkage disequilibrium in plants. **Annual Review Of Plant Biology**, Palo Alto, v. 54, n. 1, p.357-374, jun. 2003. Annual Reviews.

GEBHARDT, C. Molecular markers, maps, and population genetics. In: VREUGDENHIL, D. (Ed.). **Potato Biology and Biotechnology**: Advances and Perspectives. Amsterdam: Elsevier, 2007. Cap. 7. p. 77-89.

GEBHARDT, C.; BALLVORA, A.; WALKEMEIER, B.; OBERHAGEMANN, P.; SCHULER, K. Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. **Molecular Breeding**, Amsterdam, v. 13, n. 1, p.93-102, 2004.

GIZLICE, Z.; CARTER, T. E.; GERIG, T. M.; BURTON, J. W. Genetic Diversity Patterns in North American Public Soybean Cultivars based on Coefficient of Parentage. **Crop Science Society Of America**, Madison, v. 36, n. 3, p.753-765, 1996.

GOODMAN, M. M.. Genetic and germplasm stocks worth conserving. **Journal Of Heredity**, Washington, v. 81, n. 1, p.11-16, 1990.

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**, Berlin, v. 124, n. 3, p.447-458, 14 out. 2011.

HAO, D. R.; CHAO, M. N.; YIN, Z. T.; YU, D. Y. Genome-wide association analysis detecting significant single nucleotide polymorphisms for chlorophyll and chlorophyll fluorescence parameters in soybean (*Glycine max*) landraces. **Euphytica**, Wageningen, v. 186, n. 3, p.919-931, 2012.

HARTWIG, E. E. Varietal development. In: CALDWELL, B. E. (Ed.). **Soybeans: Improvement, Production, and Uses**. Madison: American Society Of Agronomy, 1973. p. 187-210. (Agronomy Monograph 16).

HERMANN, F. J. **A revision of the genus Glycine and its immediate allies**. Washington: United States Department Of Agriculture, 1962. 82 p. (Technical Bulletin 1268).

HIROMOTO, D. M.; VELLO, N. A.. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazilian Journal Of Genetics,** Ribeirão Preto, v. 09, n. 2, p.295-306, 1986.

HU, Z.; ZHANG, D.; ZHANG, G.; KAN, G.; HONG, D.; YU, D. Association mapping of yield-related traits and SSR markers in wild soybean (*Glycine soja* Sieb. and Zucc.). **Breeding Science**, Tokyo, v. 63, n. 5, p.441-449, 2014.

HYMOWITZ, T. On the domestication of the soybean. **Economic Botany**, NewYork, v. 24, p.408-421, 1970.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **PIB recua 0,2% e chega a R$ 1,408 trilhão no 1º trimestre de 2015**. 2015. Disponível em:

<http://saladeimprensa.ibge.gov.br/noticias?view=noticia&id=1&idnoticia=2973&busca=1&t=pib-recua-1-9-relacao-1º-tri-2015>. Acesso em: 28 ago. 2015.

IQUIRA, E.; HUMIRA, S.; FRANÇOIS, B. Association mapping of QTLs for sclerotinia stem rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach. **BMC Plant Biology**, Londres, v. 15, n. 1, p.5-16, 2015.

JUN, T.; VAN, K.; KIM, M. Y.; LEE, S.H.; WALKER, D.R. Association analysis using SSR markers to find QTL for seed protein content in soybean. **Euphytica**, Wageningen, v. 162, n. 2, p.179-191, 2007.

MACKAY, I.; POWELL, W. Methods for linkage disequilibrium mapping in crops. **Trends In Plant Science**, Oxford, v. 12, n. 2, p.57-63, 2007.

MALOSETTI, M.; VAN DER LINDEN, C. G.; VOSMAN, B.; VAN EEUWIJK, F. A. A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to *Phytophthora infestans* in Potato. **Genetics**, Austin, v. 175, n. 2, p.879-889, 1 fev. 2007.

MAPA – Ministério da Agricultura Pecuária e Abastecimento. 2015. Disponível em: <www.agricultura.gov.br/>. Acesso em Agosto/2015.

MORGANTE, M.; SALAMINI, F. From plant genomics to breeding practice. **Current Opinion In Biotechnology**, Londres, v. 14, n. 2, p.214-219, 2003.

MULATO, B. M.; MÖLLER, M.; ZUCCHI, M. I.; QUECINI, V.; PINHEIRO, J. B. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 3, p.276-283, 2010.

NASS, L. L.; VALOIS, A.C.C.; MELO, I.S.; VALADARESINGLIS, M.C. **Recursos genéticos e melhoramento Plantas**, Rondonópolis: Fundação MT, 2001. 1183p.

NELSON, R. l.; JOHNSON, E. O. C. Registration of Soybean Germplasm Lines LG97-7012, LG98-1445, and LG98-1605. **Crop Science Society Of America**, Madison, v. 46, p.1822-1824, 2006.

ORAGUZIE, N. C.; RIKKERINK, E. H. A.; GARDINER, S. E.; SILVA, H. N. de. **Association Mapping in Plants**. Nova York, Springer New York, 2007. 278 p.

PATERNIANI, E. Diversidade genética em plantas cultivadas. In: ENCONTRO SOBRE RECURSOS GENÉTICOS, 1., 1988, Jaboticabal. Anais... . Jaboticabal: UNESP/FCAVJ, 1988. p. 75 - 77.

PATTERNIANI, E.; VIEGAS, E. G. **Melhoramento e produção de milho**. 2.ed. Campinas: Fundação Cargill, 1987.

PRIOLLI, R. H. G.; MENDES-JUNIOR, C. T.; SOUSA, S. M. B.; ARANTES, N. E.; CONTEL, E. P. B. Diversidade genética da soja entre períodos e entre programas

de melhoramento no Brasil. **Pesquisa Agropecuária Brasileira,** Brasília, v. 39, n. 10, p.967-975, 2004.

PRITCHARD, J. K; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Austin, v. 155, n. 2, p.945-959, 2000.

PRITCHARD, Jonathan K.; PRZEWORSKI, Molly. Linkage Disequilibrium in Humans: Models and Data. **The American Journal Of Human Genetics**, Chicago, v. 69, n. 1, p.1-14, 2001.

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**, Washington, v. 98, n. 20, p.11479-11484, 2001.

SINGH, R. J.; NELSON, R. L.; CHUNG, G. H.. Soybean (*Glycine max* (L.) Merr.). In: SINGH, R. J. (Ed.). **Genetic Resources, Chromosome Engineering, and Crop Improvement**. Boca Raton: Crc Press, 2007b. p. 13-50. (Volume 4 Oilseed Crops).

SINGH, R J; HYMOWITZ, T. Soybean genetic resources and crop improvement. **Genome**, Ottawa, v. 42, n. 4, p.605-616, 1999.

ST. MARTIN, S. K. Effective population size for the soybean improvement program in maturity groups 00 to IV. **Crop Science Society Of America**, Madison, v. 22, p.151-152, 1982.

STUBER, C. W.; POLACCO, M.; SENIOR, M. L. Synergy of Empirical Breeding, Marker-Assisted Selection, and Genomics to Increase Crop Yield Potential. **Crop Science Society Of America**, Madison, v. 39, n. 6, p.1571-1583, 1999.

TROYER, A. F.. Selection for Early Flowering in Corn: Three Adapted Synthetics. **Crop Science Society Of America**, Madison, v. 30, n. 4, p.896-900, 1990.

WANG, J.; MCCLEAN, P. E.; LEE, K.;.GOOS, R. J; HELMS, T. Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. **Theoretical And Applied Genetics**, Berlin, v. 116, n. 6, p.777-787, 2008.

WEN, Z.; TAN, R.; YUAN, J.; BALES, C.; DU, W.; ZHANG, S.; CHILVERS, M. I.; SONG, Q.; CREGAN, P. B.; WANG, D. Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. **Bmc Genomics**, London, v. 15, n. 1, p.809-819, 2014.

WOODHAM-SMITH, C.. The Great Hunger. Londres: Penguin Books, 1962. 528 p.

WYSMIERSKI, P. T. **Contribuição genética dos ancestrais da soja às cultivares brasileira**. 99 f. Tese (Doutorado) - Curso de Genética e Melhoramento, Genética, Universidade Federal de São Paulo, Piracicaba, 2010.

2 CARACTERIZAÇÃO FENOTÍPICA DA PRODUTIVIDADE DE GRÃOS E MINERAÇÃO DE VARIÁVEIS CORRELACIONADAS EM GERMOPLASMA DE SOJA

Resumo

A seleção de plantas mais produtivas é o principal objetivo do melhoramento de plantas. Entretanto, sabe-se que devido ao uso exaustivo de cruzamentos entre materiais elites há um estreitamento da base genética das principais culturas do mundo, incluindo a soja. Assim este trabalho tem como objetivo avaliar um painel de germoplasma de soja diverso, para produtividade de grãos e variáveis correlacionadas. Para isto 95 genótipos de soja, entre eles acessos exóticos e testemunhas comerciais, foram avaliados à campo em 5 ambientes nas safras de 2012/2013 e 2013/2014 em delineamento alfa látice. As características avaliadas foram: altura da planta na maturidade (APM em cm), período de granação (PEG em dias), valor agronômico (VA em escala de notas), acamamento (AC em escala de notas), massa de cem sementes (MCS em gramas), número de dias para a maturidade (NDM), inserção da primeira vagem (IPV em cm), altura da planta no florescimento (APF em cm), teor de óleo (OLEO em %) e produtividade de grãos (PG em kg ha-1). A análise de deviance conjunta dos cinco ambientes foi realizada para todos os caracteres. Para a característica produtividade de grãos, foram calculados os BLUPs para todos os locais e os valores genéticos preditos. A partir das médias das demais características, foram calculados os coeficientes de correlação de Pearson e para aquelas variáveis com correlação significativa com a produtividade de grãos, foi realizada uma análise de árvore de regressão. Os genótipos avaliados apresentaram significância ($p<0,001$) pelo teste de LRT para todos os caracteres avaliados, assim como para a interação GXA. Os acessos exóticos Dowling, PI 331793, PI 200832, PI 170889 e PI200487 foram destaque nos ambientes analisados individualmente. Já na análise conjunta os genótipos Dowling, PI 417563, PI200526, PI 377573 e PI 159922, apresentaram valores genéticos (*u* + *g*) altos. Os caracteres que apresentaram correlação fenotípica significativa com a produtividade foram: VA, PEG, OLEO, AC e MCS. Via árvore de regressão as características que mais contribuíram em importância para produtividade de grãos foram: VA, PEG e OLEO, e portanto, devem ser consideradas para futuros estudos envolvendo a seleção de genótipos mais produtivos.

Palavras-chave: *Glycine max*; Recursos Genéticos; Interação; Melhoramento de plantas

Abstract

The selection of yield plants is the main goal of plant breeding. However it is known that due to the extensive use of crossing between materials elites, there is a narrowing of the genetic base of major crops in the world, including soybeans. So this study aims to evaluate a germplasm panel of diverse soybeans lines for grain yield and related variables. For this work, 95 soybean genotypes, including exotic access and commercial checks were evaluated in the field in five environments and at 2012/2013 and 2013/2014 seasons in alpha lattice design. The

traits evaluated were: plant height at maturity (APM in cm), grain filling period (PEG days), agronomic value (VA scale notes), lodging (AC scale notes) mass of one hundred seeds (MCS grams), number of days to maturity (NDM), first pod (IPV in cm), plant height at flowering (APF cm), oil content (OIL in%) and grain yield (PG in kg ha-1). The joint deviance analysis was performed for all the characters, for yield, the grain BLUPs to all five environments and breeding values for specific environments were calculated. From the averages of the other traits we calculated the Pearson correlation coefficients and for those variables with significant correlation with grain yield, a regression tree analysis was made. The genotypes showed significant values ($p < 0.001$) by LRT test for all traits, as well as for GXA interaction. Exotic access Dowling, PI 331793, PI 200832, PI 170889 and PI200487 were featured in the environments analyzed separately. In the joint analysis the genotypes: Dowling PI 417563, PI200526, PI 377573 and PI 159922 showed high breeding values (u + g). Traits that showed significant phenotypic correlation with productivity were: VA, PEG, OIL, AC and MCS, they were use in regression tree analysis. The features that contributed most in importance for grain yield at regression tree analysis were: VA, PEG and OIL, and therefore should be considered for future studies involving selection for yield genotype.

Keywords: *Glycine max*; Genetic resources; Interaction; Plant Breeding

2.1 Introdução

O principal objetivo de qualquer programa de melhoramento de plantas é a obtenção de cultivares mais produtivas que as atuais. Apesar dos acréscimos de produtividade ao longo dos anos, os programas de melhoramento de soja utilizaram poucos genitores nos cruzamentos, o que levou a um expressivo estreitamento da base genética da cultura nos principais países produtores (HIROMOTO e VELLO, 1986;). Essa diminuição da variabilidade pode acarretar vulnerabilidade genética e em patamares de produtividade na cultura.

Uma das estratégias para aumentar a diversidade genética dentro dos programas de melhoramento de soja e em outras culturas, tem sido a introdução de genótipos exóticos, que podem aumentar a variabilidade genética das populações de melhoramento. Em 2001/2002, uma parceria entre o USDA (Departamento de Agricultura dos Estados Unidos) e a Universidade de Illinois, lançaram 11 linhagens derivadas de cruzamentos entre acessos exóticos, dentre estas a mais produtiva originária de 4 genótipos exóticos, superou em 95% a cultivar mais produtiva dos Estados Unidos, comprovando a eficácia na utilização de germoplasma exótico (NOWLING, 2000).

Contudo, sabe-se que a subutilização dos bancos de germoplasma de soja e dos recursos genéticos desta cultura ocorre ao redor do mundo, e isso ocorre

devido principalmente à inferioridade agronômica dos acessos, comparados às cultivares comerciais, requerendo um esforço muito maior do melhorista na caracterização e avaliação destes acessos exóticos (CARTER et al., 2004).

Estas dificuldades podem ser superadas pela pré-caracterização e avaliação destes acessos exóticos, etapa denominada de pré-melhoramento (NASS et al., 2001). Entretanto, caracteres quantitativos como a produtividade de grãos estão sujeitos à influência da interação genótipo por ambiente (GxA), dificultando a repetibilidade e confiabilidade dos resultados. Assim, faz-se necessário avaliar os genótipos em múltiplos ambientes para, com isso, caracterizar os acessos mais produtivos e estáveis às regiões de interesse dos melhoristas (CARTER et al., 2004).

A caracterização de acessos exóticos pode ser complicada, por diversos fatores, dentre os principais estão: as variedades exóticas quase sempre apresentam menor produtividade do que as cultivares já adaptadas (HARTWIG e LEHMAN, 1951; WILCOX e ST.MARTIN, 1998), suscetibilidade a doenças e insetos, grupo de maturação inadequado e muitas vezes dormência de plantas. Tais atributos complicam a avaliação destes acessos em experimentos em condições de campo, resultando em parcelas com baixo estande, grande variabilidade entre diferentes ambientes e muitas vezes em parcelas perdidas, resultando consequentemente em dados experimentais desbalanceados e dificultando assim a análise de dados e validação dos resultados.

Através da utilização de modelos mistos é possível a análise de dados experimentais não ortogonais, desbalanceados e com heterogeneidade de variâncias (RESENDE, 2007), ao contrário dos normalmente exigidos, pressupostos de uma análise de variância em um modelo linear não misto. A abordagem de modelos mistos assumindo-se os efeitos de genótipos como aleatórios, possibilita a obtenção dos valores genéticos (*breeding values*) dos mesmos. Essa abordagem permite a análise de genótipos de diferentes populações e em diferentes ambientes (BERNARDO, 2010). A forma mais utilizada para predizer os componentes de variância em modelos mistos é o método da Máxima Verossimilhança Restrita (REML), proposto por Patterson e Thompson (1971), que origina predições menos viesadas para dados desbalanceados.

Além de estatísticas univariadas, a avaliação de germoplasma exótico geralmente envolve a análise de vários descritores, que muitas vezes tem pouca ou

nenhuma influência na produtividade de grãos. A fim de detectar e analisar as relações e associações entre as várias características avaliadas, analises multivariadas tais como componentes principais, análise de trilha, análise de fatores e técnicas de agrupamento são empregadas, a fim de detectar a importância das diferentes variáveis em um modelo. Além destas, a análise via árvores de regressão (BREIMAN et al., 1984), um tipo de metodologia desenvolvida para a mineração de dados multivariados (*Data mining*), também tem sido utilizada na detecção da importância das variáveis para um dado modelo. Seu uso também pode se estender para predição de modelos, tais como em análises de seleção genômica ampla.

Dado o exposto, este estudo tem como objetivo caracterizar um painel de soja, composto por acessos exóticos e padrões comerciais brasileiros, para produtividade de grãos, através do uso de modelos mistos, visando identificar genótipos exóticos mais produtivos para incorporação em um programa de melhoramento e selecionar características secundárias correlacionadas à produtividade de grãos neste mesmo painel.

2.2 Material e Métodos

2.2.1 Análise de deviance

Os ensaios foram conduzidos em cinco ambientes, sendo cada ambiente correspondente à combinação local/ano agrícola. Na safra 2012/2013, os experimentos foram conduzidos nas estações experimentais de Jaboticabal-SP, Piracicaba-SP e Ponta Grossa-PR, e na safra 2013/2014 os experimentos foram conduzidos apenas em Jaboticabal-SP e Piracicaba-SP. Foram avaliados 95 genótipos de soja, dentre eles 80 PI's (*Plant Introductions*), e 15 cultivares brasileiras (ANEXO A). Utilizou-se o delineamento experimental alfa látice 5x19 com 3 repetições, e parcelas de 4 linhas de 5 metros, com espaçamento entre linhas de 0,5 metros. Apenas as duas linhas centrais foram colhidas, evitando, assim, possíveis contaminações varietais. Os tratos culturais utilizados foram os recomendados para a região (EMBRAPA, 1999).

Os seguintes caracteres foram avaliados:

- Altura da planta no florescimento (APF) – Média das cinco plantas centrais da parcela, medidas do solo ao final da haste em cm.

- Período de granação (PEG) – Número de dias entre os estágios R5 (número de dias da semeadura até o início de enchimento de grãos) e R7 (número de dias da semeadura até a granação completa).
- Número de dias para a maturidade (NDM) – Número de dias da semeadura até 95% das vagens maduras em 50% da parcela.
- Altura da planta na maturidade (APM) – Média das cinco plantas centrais da parcela, medidas do solo ao final da haste em cm.
- Valor agronômico (VA) – Escala de notas de 1 a 5, sendo 1 para planta com baixo valor e 5 para planta excelente, para a arquitetura geral das plantas na parcela.
- Acamamento (AC) – Escala de notas de 1 a 5, sendo 5 para planta totalmente acamada e 1 para planta ereta.
- Altura da inserção da primeira vagem (IPV) – Média das cinco plantas centrais da parcela, medidas do solo até a primeira vagem da haste em cm.
- Produtividade de grãos (PROD) – Massa total das sementes produzidas na parcela, em quilograma por hectare (kg ha^{-1}).
- Massa de cem sementes (MCS) – Massa de 100 sementes em gramas (g).
- Teor de óleo (OLEO) – média em porcentagem de três leituras no espectômetro NIR (*Near-infra red spectroscopy*).

Os dados de todas as características avaliadas, foram analisados pelo software Selegen-Reml/Blup (Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos), desenvolvido por Resende e colaboradores em 1994. O modelo utilizado na análise foi o 52, utilizando modelos mistos em blocos incompletos em vários locais e uma só colheita:

$$\boldsymbol{y = Xr + Zg + Wb + Ti + e}$$

Em que:

$\boldsymbol{y}$ é o vetor das médias fenotípicas;

$\boldsymbol{r}$ é o vetor dos efeitos fixos de repetição;

$\boldsymbol{g}$ é o vetor dos efeitos aleatórios dos valores genotípicos;

$\boldsymbol{b}$ é o vetor dos efeitos aleatórios de blocos;

$\boldsymbol{i}$ é o vetor dos efeitos aleatórios da interação genótipos x ambientes;

$\boldsymbol{e}$ é o vetor dos efeitos residuais;

$\boldsymbol{X}$ é a matriz de incidência relacionada aos efeitos de repetições dentro de locais;

$\boldsymbol{Z}$ é a matriz de incidência relacionada aos valores genotipicos;

$\boldsymbol{W}$ é a matriz de incidência relacionada aos efeitos de blocos;

$\boldsymbol{T}$ é a matriz de incidência relacionada aos efeitos da interação genótipos x ambientes;

O sistema de equações matriciais que resolvem o modelo misto é:

$$\begin{bmatrix} X'X & X'Z & X'W & X'T \\ Z'X & Z'Z + A_{\lambda_1}^{-1} & Z'W & Z'T \\ W'X & W'Z & W'W + I\lambda_2 & W'T \\ T'X & T'Z & T'W & T'T + I\lambda_3 \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{g} \\ \hat{b} \\ \hat{\imath} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ T'y \end{bmatrix}$$

Em que:

$\lambda_1 = \frac{\sigma_e^2}{\sigma_g^2} = \frac{1-h^2-b^2-i^2}{h^2}$;

$\lambda_2 = \frac{\sigma_e^2}{\sigma_b^2} = \frac{1-h^2-b^2-i^2}{b^2}$;

$\lambda_3 = \frac{\sigma_e^2}{\sigma_i^2} = \frac{1-h^2-b^2-i^2}{i^2}$;

$h^2 g = \frac{\sigma_g^2}{\sigma_g^2+\sigma_b^2+\sigma_e^2+\sigma_i^2}$ (herdabilidade dos efeitos de genótipos no sentido amplo);

$b^2 = \frac{\sigma_b^2}{\sigma_g^2+\sigma_b^2+\sigma_e^2+\sigma_i^2}$ (correlação devido ao ambiente dentro de blocos);

$i^2 = \frac{\sigma_i^2}{\sigma_g^2+\sigma_b^2+\sigma_e^2+\sigma_i^2}$ (proporção da variação fenotípica explicada pela interação GxA);

$r_{gloc} = \frac{\sigma_i^2}{\sigma_g^2+\sigma_i^2}$ (correlação genotípica através dos ambientes);

σ_g^2 variância genética;

σ_b^2 variância entre blocos;

σ_i^2 variância da interação GxA;

σ_e^2 variância residual.

A significância dos efeitos do modelo foi testado pela Análise de Deviance (ANADEV), as deviances foram obtidas uma a uma, testando o modelo com e sem os efeitos de genótipos (h^2_g=0), de blocos (c2bloc=0) e da interação (c2int=0). Conforme descrito por Resende (2007). As *deviances* obtidas para cada modelo sem o referido efeito foram então subtraídas da deviance do modelo completo, os resultados foram comparados com o valor de Qui-Quadrado a 1% e 5%.

2.2.2 Árvore de regressão

O objetivo de uma árvore de regressão é explicar a variabilidade de uma variável resposta ou dependente Y em função de variáveis independentes, X, a partir de um processo de divisões binárias (FINCH e SHNEIDER, 2007), através de um processo de "poda", a árvore é construída e assim, variáveis X que tem efeitos mínimos na predição de Y são descartadas ao longo do processo. As variáveis restantes são ranqueadas de acordo com sua importância.

As médias obtidas através da análise de deviance nos 5 locais avaliados para cada característica, foram padronizadas e calculadas as correlações fenotípicas pelo Coeficiente de Correlação de Pearson no software JMP versão 12 (SAS INSTITUTE, 2015).

As médias padronizadas das variáveis com correlações significativas para produtividade de grãos foram utilizadas para a construção da árvore de regressão. Utilizando como variável resposta (Y) a produtividade de grãos (em kg ha $^{-1}$) e as variáveis correlacionadas significativamente com esta característica (VA, PEG, OLEO, AC e MCS) como variáveis independentes (X).

As divisões foram baseadas no logaritmo negativo ($-\log_{10}$(p-valor) do p-valor, associado a soma de quadrados das diferenças entre as médias de dois grupos. Para validar o melhor número de divisões da árvore de decisão, ou seja a melhor árvore, parte dos dados obtidos foram utilizados para predição do modelo e o restante para verificar a capacidade preditiva do modelo. O método de validação escolhido foi o "*K fold cross validation*" do software JMP versão 12 (SAS INSTITUTE, 2015), que divide os dados originais em *k* subgrupos, cada k é utilizado para validar o modelo feito no restante dos dados. O modelo com o melhor valor do

critério de Akaique (AIC) é então escolhido. No caso deste trabalho um valor padrão do software de *k* = 5 foi utilizado.

2.3 Resultados e Discussões

2.3.1 Análise fenotípica da produtividade de grãos

Os resultados da análise de deviance para produtividade de grãos estão sumarizados na Tabela 1. O efeito de genótipos foi considerado significativo (p<0,001) pelo teste de Qui-Quadrado para razão de verossimilhança (LTR), demonstrando que há variabilidade genética presente entre os acessos. A diversidade presente entre os genótipos, também foi corroborada pelo valor, do coeficiente de variação genotípico (CV_g = 36,01%), pois este valor possibilita a estimativa da porcentagem de variabilidade genética presente em relação a média geral.

O efeito da interação genótipo por ambientes (GXA), também apresentou valores significativos (p<0,001) pelo teste de LTR. A significância de GXA faz com que as linhagens apresentem comportamento distinto nos diferentes locais em que forem plantadas e avaliadas.

As estimativas dos demais parâmetros para produtividade de grãos também são apresentadas na Tabela 1. A herdabilidade (h^2_g) para produtividade de grãos foi igual a 0,40, indicando que 40% de variação observada corresponde à efeitos genéticos. Este valor sugere que a seleção praticada nas médias genotípicas poderá ser efetiva para a característica em questão.

Valor alto também foi observado para o coeficiente de variação (CV_e = 30%). Segundo Pimentel Gomes (1985), valores de CV inferiores a 10% são considerados baixos, valores de 10 a 20% médios, e altos quando superiores a 30%. Entretanto deve-se salientar que os genótipos utilizados, por se tratarem de indivíduos na sua maioria não adaptados e com alta diferenciação e diversidade genética, resultam em um alto valor de CV, não significando baixa precisão, mas sim grande variabilidade genética.

De acordo com Resende (2007), o coeficiente de variação normalmente utilizado para avaliar a precisão experimental não é indicado para isso. Segundo o autor, a medida mais indicada seria a acurácia, a refere-se à correlação entre o valor genotípico verdadeiro e aquele estimado a partir dos dados obtidos em campo.

Para o progresso em programas de melhoramento, devem ser almejados valores de acurácia acima de 70% (Resende, 2007). Neste trabalho a estimativa de acurácia obtida foi de 92%, confirmando a boa precisão experimental dos ensaios.

Os coeficientes de determinação de genótipos (h2g), de blocos (c2bloc) e da interação GXA (c2int) ainda na Tabela 1, representam o quanto da variação fenotípica é explicada por cada um desses componentes. Sendo assim os efeitos genotípicos contribuíram com 40,6% da variação, blocos representou pouco mais de 0,3% e a interação genótipos por ambientes 24,6% da variação total.

Tabela 1- Resumo da análise de deviance de 95 genótipos de soja avaliados em 5 ambientes

Efeitos	**Deviance**	**LTR Qui-Quadrado**	**Componentes de Variância**	**Coef.Deter**
Genótipos	19524,21	$167,75^{**}$	401278,19	h2g=0,4066
Blocos	19356,60	$0,14^{ns}$	3245,21	c2bloc=0,0032
Genótiposxlocais	19538,62	182^{**}	243348,53	c2int=0,2466
Resíduo	-	-	338972,55	-
Modelo Completo	19356,46	-	-	-
h^2_g (Herdabilidade ajustada de parcelas individuais)				0,40
Ac_{gen} (Acurácia da seleção de genótipos)				0,92
CV_g% (Coeficiente de variação genotípico)				36,01
CV_e% (Coeficiente de variação residual)				33,09
Média Geral (kg ha^{-1})				1759

*Qui – quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5% e 1%, respectivamente

A classificação dos melhores 20 genótipos para a característica produtividade de grãos, para os cinco ambientes individuais, e na análise conjunta de todos os locais com os valores de BLUP desconsiderando a interação GXA, está apresentada na Tabela 2.

Os acessos selecionados com os melhores valores genéticos preditos (*u+g*), na análise conjunta foram as linhagens, PI 417563, Dowling, PI 200526, PI 377573 e PI 159922, ficando abaixo apenas das variedades brasileiras usadas como testemunhas. Estes acessos exóticos apresentaram cerca de 1,25% a mais de produtividade em relação a média geral (1759 kg ha^{-1}) do painel.

Segundo Resende (2004), estes valores genéticos preditos (*u+g*) podem ser utilizados para orientar a recomendação de linhagens para diferentes regiões, com valores de GXA diferenciais, ou até mesmo locais com alta heterogeneidade ambiental, visto que é esperado que estes genótipos apresentem essa mesma produtividade nos diferentes ambientes.

Analisando as características destes cinco genótipos exóticos que se destacaram na análise conjunta, observam-se caracteres interessantes ao melhoramento. A PI417563 foi coletada no Vietnã, possui hábito indeterminado, grupo de maturação VI e resistência ao Cancro da Haste (*Diaporthe phaseolorum var. caulivora*). Já a variedade Dowling, é uma cultivar plantada nos Estados Unidos, desenvolvida no Texas em meados da década de 70, pelo cruzamento entre as linhagens Semmes e Komata. Apresenta grupo de maturação VIII, hábito de crescimento determinado e características de resistência interessantes como, resistência a Podridão Radicular de Fitóftora (*Phytophthora sojae*), Pústula Bacteriana (*Xanthomonas campestris pv. Phaseoli*) e ao Pulgão da Soja (*Aphis glycines Matsumura*). Coletado no Japão o acesso PI200526, possui hábito determinado, resistência ao Cancro da Haste e grupo de maturação VIII. Uma das características diferenciais desta linhagem é a presença de resistência tolerância a estresse salino. A PI 377573, por sua vez, possui resistência a uma gama de doenças, tais como Cancro da Haste, Podridão de Fitóftora e Pústula Bacteriana, originária da China este acesso apresenta hábito determinado e grupo de maturação VII, seu destaque vem pela moderada resistência ao Nematóide de Cisto raça 3. Convergindo para uma região mais próxima ao Brasil, encontra-se a última linhagem a PI 159922, coletada no Peru, tem como características:, hábito indeterminado, grupo de maturação VIII e moderada resistência a Nematoide do Cisto Raça 5. É importante destacar que todos os cinco acessos possuem grupos de maturação próximos aos utilizados nas regiões de plantio brasileiras, sendo este o provável motivo de sua boa adaptação (USDA, 2013).

Já quando se capitaliza os efeitos da interação ($u + g + ge$) mostrados na Tabela 2, a recomendação deve ser feita para cada local. No ambiente 1, por exemplo, o genótipo exótico mais indicado para a incorporação em um programa de melhoramento seria a variedade Dowling, já descrita anteriormente. Já no ambiente 2, se destaca a PI 331793, superior a umas das testemunhas comerciais mais plantadas no Sul do país a BMX Potência. Este acesso do Vietnã de hábito indeterminado é resistente a Podridão Radicular e grupo de maturação VIII. O acesso PI 200832 por sua vez, foi o acesso exótico destaque no ambiente 3, superando as testemunhas comerciais IAC 100 e Paranagoiania, dentre os caracteres peculiares pode-se ressaltar, a resistência para Cancro da Haste e o grupo de maturação VIII. Superando duas testemunhas, A7002 e Pintado, no

ambiente quatro temos a PI 170889, originária da África do Sul e com grupo de maturação VI. E por fim no ambiente 5, superando as cultivares comerciais Conquista e IAC 100, o acesso PI 200487 do Japão, também resistente ao Cancro da Haste e grupo de maturação VIII (USDA, 2013).

Estudos envolvendo germoplasma de soja e a contabilização dos efeitos da interação GXA são escassos na literatura, o que pode ser consequência da dificuldade encontrada na condução de tais experimentos principalmente quando em mais de um local. Os principais trabalhos desenvolvidos são principalmente ligados a resistência a pragas e doenças (Miles et al., 2006; Niide et al., 2012; Pathan et al., 2014) visando a incorporação de alelos de resistência.

Em trabalho iniciado por Mulato (2010), Sigrist (2012) caracterizou 81 acessos, pertencentes ao mesmo painel utilizado neste mesmo estudo, via marcadores agro morfológicos em uma safra de experimento a campo. Apesar de utilizar apenas um local em sua avaliação o autor encontrou resultados similares aos apresentados aqui, com ampla variabilidade entre os acessos e coeficiente de variação alto para produtividade de grãos.

Tabela 2 – Valores genotípicos estimados para característica produtividade de grãos (kg ha^{-1}), para os 20 melhores genótipos de soja em 5 ambientes e para a análise conjunta

Ambiente 1		Ambiente 2		Ambiente 3		Ambiente 4		Ambiente 5		Conjunta	
Genótipos	***u+g+ge***	**Genótipos**	***u+g+ge***	**Genótipos**	***u+g+ge***	**Genótipos**	***u+g+ge***	**Genótipos**	***u+g+ge***	**Genótipos**	***u+g***
Potência	3995,76	Paranagoiana	4067,90	Potência	3297,44	LQ 1413	3613,34	Potência	4011,59	Potência	3332,64
CD215	3650,43	CD215	4046,03	CD215	2911,88	Potência	3607,24	CD215	3268,47	CD215	3307,52
LQ 1505	3564,66	LQ 1421	3945,08	VMáx	2675,43	CD215	3599,79	VMáx	3057,18	VMáx	2948,67
JAB 00-02	3548,98	LQ 1505	3903,09	PI 200832	2485,50	LQ 1050	3376,59	LQ 1413	2975,48	Paranagoiana	2905,46
LQ 1050	3524,15	Conquista	3824,52	PI 417581	2337,09	Conquista	3361,91	A7002	2902,36	LQ 1505	2864,15
Paranagoiana	3518,34	JAB 00-02	3781,83	PI 148260	2085,48	LQ 1505	3259,18	Sambaíba	2895,97	JAB 00-02	2860,04
VMáx	3422,54	Pintado	3710,44	Dowling	2083,47	LQ 1421	3235,69	Paranagoiana	2846,00	LQ 1413	2845,65
Conquista	3333,89	IAC100	3663,09	JAB 00-05	2026,79	Sambaíba	3139,43	LQ 1050	2749,49	LQ 1050	2819,96
LQ 1421	3199,69	LQ 1413	3577,92	Paranagoiana	1988,94	IAC100	3043,45	JAB 00-02	2726,94	Conquista	2815,22
LQ 1413	3179,17	VMáx	3500,30	PI 84910	1979,24	JAB 00-02	3041,24	Kinoshita	2620,12	IAC100	2762,80
IAC100	3160,31	JAB 00-05	3355,21	IAC100	1977,62	JAB 00-05	2999,40	PI 210352	2615,22	LQ 1421	2739,21
Pintado	3137,75	LQ 1050	3299,07	PI 322695	1974,49	VMáx	2809,32	Shira Nuhi	2590,75	JAB 00-05	2592,03
Dowling	3108,29	Sambaíba	3152,90	PI 153681	1941,55	Paranagoiana	2801,32	IAC100	2578,23	Sambaíba	2521,43
PI 360851	2903,04	PI 331793	2845,28	PI 417582	1920,94	PI 170889	2794,75	LQ 1505	2573,10	Pintado	2418,62
PI 417563	2895,55	PI 377573	2758,61	PI 145079	1882,23	Dowling	2679,39	Conquista	2547,10	Dowling	2341,23
Sambaíba	2874,67	Potência	2705,40	PI 90577	1871,56	Pintado	2568,95	PI 417563	2540,75	PI 417563	2327,26
JAB 00-05	2832,42	PI 274454-A	2690,26	JAB 00-02	1868,86	PI 210352	2541,45	PI 159922	2393,27	Shira Nuhi	2153,63
PI 377573	2829,73	PI 259540	2642,21	PI 253664	1867,23	PI 331795	2499,81	JAB 00-05	2251,46	A7002	2141,29
PI 159922	2778,31	PI 159922	2574,15	PI 205384	1850,42	A7002	2468,00	PI 210178	2201,33	PI 377573	2114,13
PI 79861	2656,27	PI 416828	2550,48	PI 36906	1837,12	PI 417563	2392,82	LQ 1421	2187,10	PI 159922	2106,11

*Ambiente 1- Piracicaba 2012/2013, ambiente 2- Jaboticabal 2012/2013, ambiente 3 - PontaGrossa 2012/2013, ambiente 4- Piracicaba 2013/2014, ambiente 5- Jaboticaba l 2013/2014. **O somatório de *u+g+ge*, equivale a média (*u*) do local *j*, somada aos efeitos de genótipos (*g*) e da interação genótipos por ambientes(*ge*)

Alguns trabalhos envolvendo os genótipos exóticos destacados pela análise de deviance podem ser encontrados na literatura. Para a cultivar Dowling, vários estudos testando a resistência a insetos têm sido realizados, devido a sua resistência por antibiose ao Pulgão da Soja (HILL et al., 2004). Laumann et al. (2008), avaliou à campo diferentes genótipos de soja, dentre estes a cultivar Dowling, sob infestação natural de percevejos. Segundo os autores os níveis de infestação das populações de percevejo em Dowling foram bem inferiores dos demais genótipos, mesmo quando comparados a cultivar resistente a percevejos IAC 100.

Para a PI 200526, também denominada Shira Nuhi, estudos relacionados com a Ferrugem Asiática (*Phakopsora pachyrhizi*) são encontrados, visto que um dos genes de resistência a doença (Rpp5) foi descrito como presente na mesma (GARCIA et al., 2008). Pierozzi et al., (2008), utilizando 11 genótipos dentre eles a PI 200526 e 55 linhagens derivadas de cruzamentos bialélicos entre as mesmas 11 linhas, avaliou à campo o tipo de lesão da Ferrugem. Os autores obtiveram resultados indicando que a as PI's 200526 e 200487 carregam genes de resistência de efeitos maiores, os quais são diferentes dos genes *Rpp2* e *Rpp4*.. Assim como a variedade Shira Nuhi, a PI200487, também denominada Kinoshita, está em diversos trabalhos relacionados com resistência a Ferrugem Asiática, envolvendo o mesmo gene de resistência (Rpp5) descrito anteriormente.

2.3.2 Mineração de variáveis via árvore de regressão

A partir dos resultados da análise de deviance todos os caracteres agronômicos analisados apresentaram efeitos significativos ($p<0,001$) pelo teste da razão de verossimilhança (LTR) (ANEXO B). As seguintes características apresentaram correlações fenotípicas (Tabela 3) significativas com a produtividade de grãos pelo teste t: PEG (0,487), VA (0,718), AC (-0,438), MCS (0,244), e OLEO (0,530). Em contrapartida, os caracteres relacionados a arquitetura de planta tais como APM, APF e IPV não foram significativamente associados ao rendimento de grãos, sendo assim não utilizados na composição da árvore de regressão.

A característica VA também apresentou valores de correlações fenotípicas significativas para a maioria dos caracteres avaliados (Tabela 3), ilustrando sua

utilização como avaliador de características agronômicas favoráveis na planta, aumentando positivamente caracteres como PEG, MCS e OLEO e diminuindo características desfavoráveis como APM, AC e APF.

As correlações entre OLEO e os demais caracteres foram todas significativas, sendo ela favorável para os caracteres relacionados à produtividade de grãos, PEG, VA e MCS. E negativa para os caracteres APM, AC, APF, IPV e NDM. Além disso, a literatura tem frequentemente relatado em diversos estudos a correlação negativa entre OLEO e teor de proteína em soja (THORNE e FEHR, 1970; HARTWIG e HINSON, 1972; HYMOWITZ et al., 1972; SHANNON et al., 1972; VOLDENG et al., 1997; WILCOX & GUODONG, 1997).

Para a característica teor de óleo segundo trabalho realizado por Abdelnor, (1995), no inicio do plantio desta oleaginosa no Brasil as seleções de plantas foram feitas para dois caracteres principais, a produtividade de grãos e o teor de óleo, fazendo com que as cultivares brasileiras tenham possivelmente valores correlacionados entre estes dois caracteres.

Para PEG há significâncias das correlações entre todos os caracteres, exceto para IPV, sendo positivas as correlações para as caracteristicas relacionados a produtividade de grãos já citados anteriormente, assim como também para o caráter NDM, visto que, assim como PEG é uma característica relacionada ao número de dias gasto pela planta. Já para as características negativamente correlacionadas tem-se APM, AC e APF, sendo estas todas relacionadas a arquitetura da planta.

Estudando as correlações fenotípicas entre os caracteres em dois métodos de condução de populações segregantes em soja, Rocha et al. (2015), detectou correlações significativa entre produção de grão e PEG, tanto no método genealógico como na metodologia de descendente de uma única semente. Entretanto, Bermudez (2015) avaliando duas populações originárias de cruzamento com duas PI`s, detectou a não correlação entre PEG e produção de grãos nas duas populações analisadas. Portanto, a correlação entre este dois caracteres deve ser melhor estudada para utilização como caráter na seleção para produtividade de grãos.

Visando melhorar o entendimento sobre as relações entre os caracteres agronômicos e a produtividade grãos, realizou-se uma mineração das variáveis (*Data*

mining) via análise de regressão, utilizando as variáveis fenotipicamente correlacionadas com a produtividade de grãos (PEG, VA, AC, MCS, OLEO). O modelo explicando a maior quantidade de variação na produtividade de grãos foi uma árvore de regressão, com menor valor de AIC, contendo doze nós terminais (Figura 1).

Tabela 3 – Coeficientes de correlações fenotípicas de Pearson entre os caracteres altura da planta na maturidade (APM em cm), período de granação (PEG em dias), valor agronômico (VA em escala de notas), acamamento (AC em escala de notas), massa de cem sementes (MCS em gramas), número de dias para a maturidade (NDM), inserção da primeira vagem (IPV em cm), altura da planta no florescimento (APF em cm), teor de óleo (OLEO em %), produtividade de grãos (PROD em kg ha-1), considerando a média dos 5 ambientes

	APM	**PEG**	**VA**	**AC**	**MCS**	**NDM**	**IPV**	**APF**	**OLEO**	**PROD**
APM		-0,224*	-0,347*	0,731*	-0,563*	0,606*	0,736*	0,928*	-0,561*	-0,056ns
PEG			0,467*	-0,519*	0,464*	0,256*	-0,162ns	-0,243*	0,538*	0,487*
VA				-0,753*	0,540*	-0,171ns	-0,083ns	-0,211*	0,646*	0,718*
AC					-0,683*	0,347*	0,454*	0,630*	-0,717*	-0,438*
MCS						-0,449*	-0,283*	-0,566*	0,614*	0,244*
NDM							0,454*	0,615*	-0,307*	0,091ns
IPV								0,741*	-0,373*	-0,080ns
APF									-0,516*	0,005ns
OLEO										0,530*
PROD										

** Significativo a 1% de probabilidade pelo teste t

ns Não significativo a 1% e 5% de probabilidade pelo teste t

A análise por árvore de regressão demonstrou que as variáveis VA, PEG e OLEO são responsáveis pela maior parte da variabilidade dentre as variáveis avaliadas para produtividade de grãos (Tabela 4), correspondendo aos melhores preditores a serem utilizados quando o objetivo é o aumento da média para a referida característica.

A primeira divisão da árvore ocorreu pelo caráter VA, indicando que esta foi a variável avaliada que mais influenciou no rendimento de grãos, contribuindo com 73,03% da variação total. Seguida respectivamente pelos caracteres PEG (13,23%) e OLEO (8,81%), contabilizando 95% da variação total observada (Tabela 4).

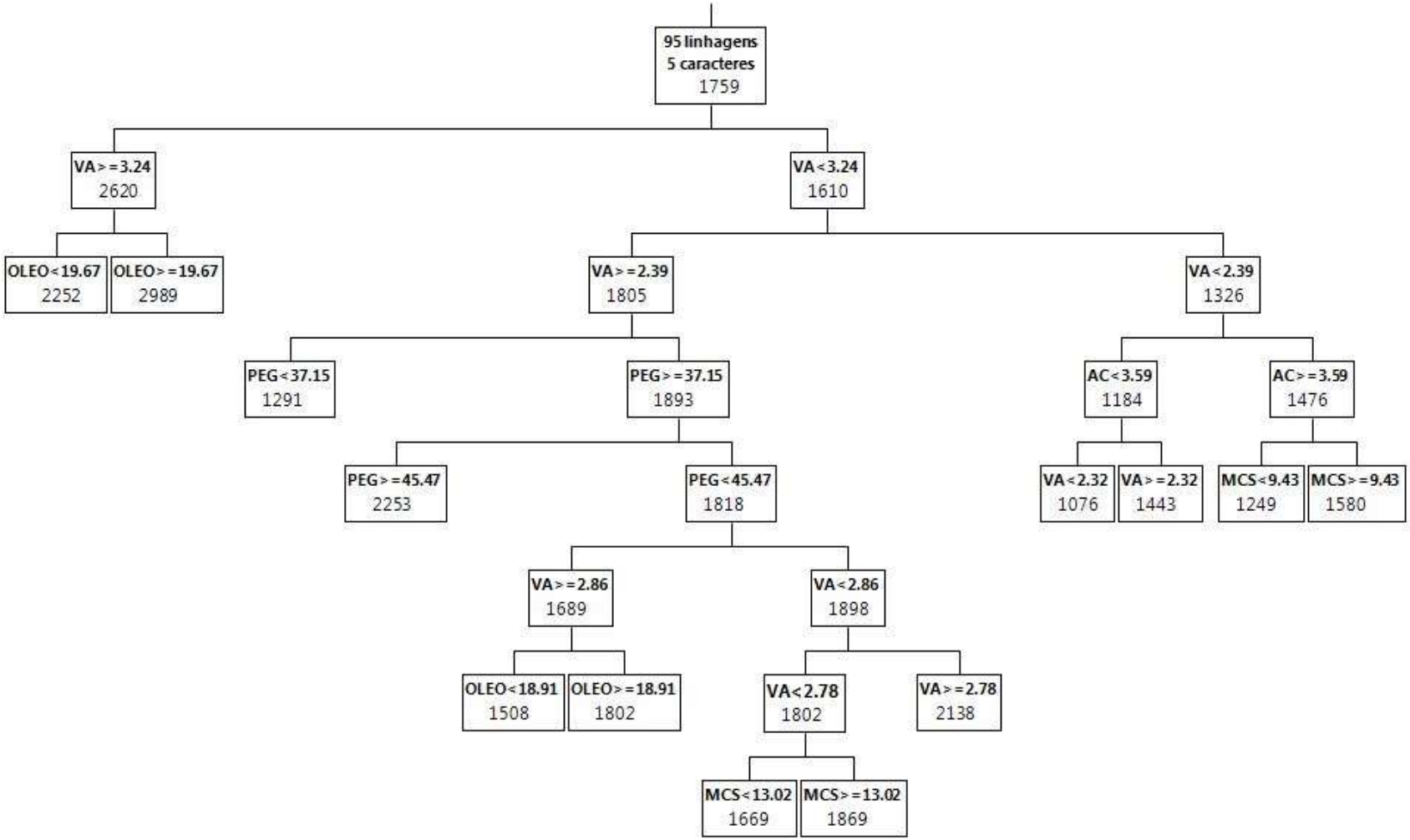

Figura 1 – Árvore de regressão da produtividade de grãos (kg ha $^{-1}$) em soja predita por caracteres agronômicos, período de granação (PEG em dias), valor agronômico (VA em escala de notas), acamamento (AC em escala de notas), massa de cem sementes (MCS em gramas), teor de óleo (OLEO em %)

Para VA, foram realizadas cinco divisões (Figura 1), o grupo formado com maior média de produtividade (2989 kg ha^{-1}), foi aquele com VA maior que 3,24 e teor de óleo superior a 19,679. Sete indivíduos atenderam a estes requisitos, sendo eles: CD 215, Conquista, Vmáx, BMX Potência, JAB 00-02-2/763D, LQ 1050 e LQ 1413, dentre estes constam os padrões comerciais plantados no Brasil utilizados como testemunhas.

O segundo grupo com maior média em produtividade de grãos (2253 Kg ha^{-1}) foi formado pelos genótipos com notas de VA menor que 3,24 e maior que 2,39 e PEG maior que 37,15 e maior que 45,47, sete genótipos possuíam tais características os quais foram: PI 145079, PI59097, PI 170889, Paranagoiania, A7002, Sambaíba e Dowling. O terceiro grupo com maior média produtividade (2252 kg ha^{-1}) foi formado assim como o primeiro, com notas de VA maiores que 3,24 e porcentagem de teor de óleo menor que 19,67, sendo pertencentes a este grupo as linhagens: PI 341264, PI 381660, PI407764, IAC 100, JAB 00-05-6/763D, LQ1505 e LQ1421.

Em contrapartida, o grupo com menor média de produtividade de grãos (1076 kg ha^{-1}) foram os genótipos com notas de VA menores que 2,39, AC menor que 3,59 e VA menor que 2,32.. A média baixa deste grupo pode ser explicada facilmente, visto que, valores baixos de VA indicam baixo valor agronômico enquanto que valores altos de AC indicam plantas acamadas o que prejudica o potencial agronômico das plantas.

Outra divisão com moderada produtividade de grãos (2139 kg ha^{-1}) e composto quase totalmente por acessos exóticos, exceto pela cultivar Pintado, contém linhagens com notas de PEG menores que 45,47 dias, VA menor do 2,86 e maior do que 2,78, foram: PI 285095, PI377573, PI148260, PI417563 e a PI200487 (Kinoshita).

É importante salientar que, além de produtivos as melhores linhagens e acessos classificados pelos caracteres agronômicos possuem características importantes como o bom valor agronômico, o qual engloba vários caracteres simultaneamente, boa arquitetura de planta, teor de óleo e período de enchimento de grãos adequado aos padrões estabelecidos para a cultura.

Tabela 4 – Proporção da variabilidade explicada pelos caracteres agronômicos, valor agronômico (VA em escala de notas), período de granação (PEG em dias), teor de óleo (OLEO em %), acamamento (AC em escala de notas) e massa de cem sementes (MCS em gramas)

Variáveis	Número de Divisões	Proporção (%)
VA	5	73,03
PEG	2	13,23
OLEO	2	8,81
AC	1	2,86
MCS	2	2,07

Estudos envolvendo árvores de regressão já foram utilizados por outros autores para predição da variabilidade da produtividade de grãos na agricultura em diversas culturas (Lobell et al., 2005; Tittonell et al., 2008; Zheng et al., 2009). Em soja, Zheng et al. (2010), utilizou a metodologia para detectar quais parâmetros ligados ao solo, estão envolvidos na variabilidade da produtividade de soja sob estresse hídrico. Segundo os autores os principais componentes responsáveis pela variação no rendimento foram os teores de fósforo disponíveis e aplicados.

Apesar de trabalhos utilizando árvores de regressão na seleção de caracteres em soja serem relativamente novos na literatura, trabalhos analisando a correlação entre produtividade de grãos e caracteres agronômicos na cultura da soja, foram amplamente reportados ao longo dos anos (MORO et al., 1992; YOKOMIZO et al., 2000; Fehr et al., 2003; PANTHEE et al., 2005).

Em trabalho realizado por Bárbaro et al. (2007), os autores avaliaram as correlações, análise de trilha e correlação em sete populações de soja derivadas de cruzamentos, dentre os resultados foram detectadas correlações positivas e significativas entre o caractere VA e PROD em todas as populações analisadas, corroborando a tendência de que plantas com valores maiores de VA serem mais produtivas. Lopes et al., (2002), também encontrou correlação positiva e significativa entre PROD e VA em populações de soja F_2.

2.4 Conclusões

Os acessos exóticos Dowling, PI 417563, PI200526, PI 377573 e PI 159922, apresentaram boa produtividade de grãos nos cincos ambientes avaliados neste

trabalho e podem ser indicadas para incorporação em um programa de melhoramento para as regiões de São Paulo e Paraná.

As linhagens "*LQ's*" do programa de melhormento de soja da Escola Superior de Agricultura "Luiz de Queiroz" e "JAB" do programa de melhoramento da Universidade Estadual Paulista "Júlio de Mesquita" – Campus Jaboticabal, apresentaram ótimo desempenho, sendo ranqueadas entre as 20 melhores linhagens na maioria dos ambientes avaliados. Portanto também devem ser recomendadas para inclusão em programas de melhoramento de soja.

As características VA e OLEO possuem correlação positiva e significativa com a produtividade de grãos e podem ser consideradas na seleção de genótipos com média superior para rendimento de grãos. Mais estudos devem ser conduzidos afim de entender a relação entre PEG e produção de grãos.

A característica AC está negativamente correlacionada com a produtividade de grãos e a seleção para diminuir este caráter provocará um acréscimo na produtividade média das variedades.

Referências

BÁRBARO, I. M.; CENTURION, M. A. P. C.; DI MAURO, A. O.; UNÊDA-TREVISOLI, S. H.; COSTA, M. M. Comparação de estratégia de seleção no melhoramento de populações F5 de soja. **Ceres**, Viçosa, v.54, n.313, p.250-261, 2007.

BERNARDO, R. **Breeding for quantitative traits in plants**. Minesota: Stema Press, 2002. 369 p.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE C. J. **Classification and Regression Trees**. Nova York: Chapman And Hall, 1984. 254 p.

FEHR, W. R.; HOECK, J. A.; JOHNSON, S. L.; MURPHY, P.A.; NOTT, J.D.; PADILLA, G.I.; WELKE, G.A. Genotype and Environment Influence on Protein Components of Soybean. **Crop Science Society Of America**, Madison, v. 43, n. 0, p.511-514, 2003.

FINCH, H.; SCHNEIDER, M. K. Classification Accuracy of Neural Networks vs. Discriminant Analysis, Logistic Regression, and Classification and Regression Trees. **Methodology**, Boston, v. 3, n. 2, p.47-57, 2007.

GARCIA, A.; CALVO, E.S.; KIIHL, R.A.S.; HARADA, A.; HIROMOTO, D.M.; VIEIRA, L.G. Molecular mapping of soybean rust (*Phakopsora pachyrhizi*) resistance genes:

discovery of a novel locus and alleles. **Theoretical and Applied Genetics**, v.117, p. 545-553, 2008.

HARTWIG, E.E.; HINSON, K. Association between chemical composition of seed and seed yield of soybeans. **Crop Science Society Of America**, Madison, v.12, p.829-830, 1972.

HARTWIG, E.e.; LEHMAN, S.g.. Inheritance of resistance to the bacterial pustule disease in soybean. **Agronomy Journal**, Madison, v. 43, p.226-229, 1951.

HILL, C. B.; LI, Y.; HARTMAN, G. L. Resistance to the soybean aphid in soybean germplasm. **Crop Science Society Of America**, Madison, v. 44, p.98-106, 2004.

HIROMOTO, D. M.; VELLO, N. A.. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazilian Journal Of Genetics,** Ribeirão Preto, v. 09, n. 2, p.295-306, 1986.

HYMOWITZ, T.; COLLINS, F. I.; PANCZNER, J.; WALKER, W. M. Relationship between the content of oil, protein, and sugar in soybean seed. **Agronomy Journal**, Madison, v.64, p.613-616, 1972.

KWON, S.H.; TORRIE, J.H. Heritability of and interrelationships among traits of two soybean populations. **Crop Science Society Of America**, Madison, v. 04, p.196-198, 1964.

LAUMANN R. A.; FARIAS NETO, A. L.; BLASSIOLI-MORAES, M. C.; SILVA, A. P.; VIEIRA, C. R.; MORAES, S. V. P.; HOFFMAN-CAMPO, C. B.; BORGES, M. Dinâmica populacional de percevejos (Hemiptera:Pentatomidae) em diferentes genótipos de soja. In: IX Simpósio Nacional Cerrado; II Simpósio Internacional Savanas Tropicais, Brasília, DF. **Desafios e estratégias para o equilíbrio entre sociedade, agronegócio e recursos naturais: anais**. Planaltina: Embrapa Cerrados. 2008.

LOBELL D. B.; ORTIZ-MONASTERIO J. I.; ASNER G. P.; NAYLOR R. L.; FALCON W. P. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. **Agronomy Journal**, Madison, v. 97, p.241-249, 2007.

LOPES, A. C. DE A.; VELLO, N. A.; PANDINI, F.; ROCHA, M. DE M.; TSUTSUMI, C. Y. Variabilidade e correlações entre caracteres em cruzamentos de soja. **Scientia Agrícola**, Piracicaba, v. 59, p.341-348, 2002.

MILES, M. R.; FREDERICK, R. D.; HARTMAN, G. L.. Evaluation of Soybean Germplasm for Resistance to *Phakopsora pachyrhizi*. **Plant Health Progress**, St. Paul, v. 10, p.1-29, 2006

MORO, G. L.; REIS, M. S.; SEDIYAMA, C. S.; SEDIYAMA, T.; OLIVEIRA, A. B. Correlações entre alguns caracteres agronômicos em soja (*Glycine max* (L.) Merrill). **Revista Ceres**, Viçosa, v.39, n.223, p.225-232, 1992.

MULATO, B. M.; MÖLLER, M.; ZUCCHI, M. I.; QUECINI, V.; PINHEIRO, J. B. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. **Pesquisa Agropecuária Brasileira**., v. 45, n. 3, p. 276-283, 2010.

NASS, L. L.; MIRANDA-FILHO, J. B.; SANTOS, M. X.. Uso de germoplasma exótico no melhoramento. In: NASS, L. L.; VALOIS, A. C. C.; MELO, I. S.; VALADARES-INGLIS, M. C. (Ed.). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação Mato Grosso, 2001. p. 101-122.

NIIDE, T.; HIGGINS, R. A.; WHITWORTH, R. J.; SCHAPAUGH, W. T.; SMITH, C. M.; BUSCHMAN, L. L. Antibiosis Resistance in Soybean Plant Introductions to *Dectes texanus* (Coleoptera: Cerambycidae). **Journal Of Economic Entomology**, Lanham, v. 105, n. 2, p.598-607, 2012.

NOWLING, G. L. The uniform soybean tests, northern region 2000. Departament of Agronomy- Purdue University, **USDA-ARS**, West Lafayette, 2000.

PANTHEE, D.R.; PANTALONE, V.R.; WEST, D.R.; SAXTON, A.M.; SAMS, C.E. Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. **Crop Science Society Of America**, Madison, v.45, p.2015-2022, 2005.

PATHAN, S. M.; LEE, J. D.; SLEPER, D. A.; FRITSCHI, F. B.; SHARP, R. E.; CARTER, T. E.; NELSON, R. L.; KING, C. A.; SCHAPAUGH, W. T.; ELLERSIECK, M. R.; NGUYEN, H. T.; SHANNON, J. G. Two Soybean Plant Introductions Display Slow Leaf Wilting and Reduced Yield Loss under Drought. **Journal of Agronomy and Crop Science**, Madison, v. 200, p.231–236. 2014.

PATTERSON, H.D.; THOMPSON, R. Recovery of inter-block information when blocks sizes are unequal. **Biometrika**, Oxford, v.58, p.545-554, 1971.

PIEROZZI, P. H. B.; RIBEIRO, A. S.; MOREIRA, J. U. V.; LAPERUTA, L. C.; RACHID, B. F.; LIMA, W. F.; ARIAS, C. A. A.; OLIVEIRA, M. F.; TOLEDO, J. F. F. DE. New soybean (*Glycine max*, Fabales, Fabaceae) sources of qualitative genetic resistance to Asian soybean rust caused by *Phakopsora pachyrhizi* (Uredinales, Phakopsoraceae). **Genetics and Molecular Biology**, Ribeirão Preto, v.31, p.505-511, 2008.

PIMENTEL GOMES, F. **Curso de Estatística Experimental**. São Paulo: Nobel, 467 p, 1985.

RESENDE, M. D. V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007.

RESENDE, M. D. V. **Métodos estatísticos ótimos na análise de experimentos de campo**. Colombo: Embrapa Floresta, 2004. (Documentos, 100).

ROCHA, F DA. **Seleção de genótipos de soja para resistência ao complexo de percevejos**. 2015. 81p. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2015.

SHANNON, J. G.; WILCOX, J. R.; PROBST, A. M. Estimated gains from selection for protein and yield in the F4 generation of six soybean populations. **Journal of Agronomy and Crop Science**, Madison, v.12, p.824-826, 1972.

SIGRIST, M. S. **Mapeamento associativo de locos relacionados à produtividade de grãos em soja**. 81p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2012.

THORNE, J. C.; FEHR, W. R. Incorporation of highprotein, exotic germplasm into soybean populations by 2- and 3-way crosses. **Crop Science Society Of America**, Madison, v.10, p.652-655, 1970.

TITTONELL, P.; SHEPHERD, K. D.; VANLAUWE, B.; GILLER, K. E.; Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—An application of classification and regression tree analysis. **Agriculture, Ecosystems & Environment**, Amsterdam, v. 123, n. 1-3, p.137-150, 2008.

USDA - UNITED STATES DEPARTMENT OF AGRICULTURE. . GRIN - **Germplasm Resources Information Network. National Germplasm Resources Laboratory**. 2015. Disponível em: <http://www.ars-grin.gov>. Acesso em: 26 maio 2015.

VOLDENG, H.D.; COBER, E.R.; HUME, D.J.; GILLARD, C.; MORRISON, M.J. Fifty-eight years of genetic improvement of short-season soybean cultivars in Canada. **Crop Science Society Of America**, Madison, v.37, p.428-431, 1997.

WILCOX, J.R.; GUODONG, Z. Relationship between seed yield and seed protein in determinate and indeterminate soybean populations. **Crop Science Society Of America**, Madison, v.37, p.361-364, 1997.

WILCOX, J. R.; ST. MARTIN, S. K. Soybean Genotypes Resistant to Phytophthora sojae and Compensation for Yield Losses of Susceptible Isolines. **Plant Disease**, St. Paul, v. 82, n. 3, p.303-306, 1998.

YOKOMIZO, G. K.; DUARTE, J. B.; VELLO, N. A. Correlações fenotípicas entre tamanho de grãos e outros caracteres em topocruzamentos de soja tipo alimento com tipo grão. **Pesquisa Agropecuária Brasileira**, Brasília, v. 35, p. 2235-2241, 2000.

ZHENG H. F.; CHEN L. D.; HAN X. Z.; ZHAO X. F.; MA Y. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. **Agriculture, Ecosystems & Environment**, Amsterdam, v. 132, n. 1-2, p.98-105, 2009.

3 ANÁLISE FENOTÍPICA DA DIVERSIDADE GENÉTICA EM PAINEL DE ACESSOS DE SOJA.

Resumo

A base genética da soja cultivada no Brasil é estreita, tal perda de variabilidade pode acarretar em patamares de produtividade e vulnerabilidade genética na cultura. Para o aumento da base genética é necessário a incorporação de novas fontes de variabilidade como acessos de um banco de germoplasma, uma das estratégias para incorporação de linhagens exóticas é caracterização fenotípica de tais genótipos. Neste trabalho foi realizada a caracterização fenotípica de 80 acessos exóticos de soja e 15 testemunhas comerciais brasileiras. Dez características agro morfológicas foram avaliadas em 4 experimentos na safras 2012/2013 e 2013/2014. Foram realizadas análises de Deviance, Componentes Principais e dois métodos de agrupamento: Ward's e Average Linkage. Todas as variáveis apresentaram significância para a análise de deviance nos genótipos avaliados, indicando boa variabilidade genética entre os mesmos. Grande parte de variância total apresentada foi explicada por 3 componentes principais, entretanto os dois primeiros componentes permitiram maior poder de diferenciação dos genótipos. No agrupamento de *Ward's* foram formados dez grupos e no, no método de *Average Linkage* as linhagens foram divididas em sete grupos. Com base nos resultados foi possível concluir que, os acessos deste painel de soja possuem divergência genética, com base em marcadores agro morfológicos, sendo assim, é possível a seleção de genótipos para utilização em programas de melhoramento de soja visando o aumento da variabilidade genética.

Palavras-chave: *Glycine max*; Agrupamento; Germoplasma exótico

Abstract

The genetic basis of soybeans grown in Brazil is narrow, such a loss of variability can result in productivity levels and genetic vulnerability in the culture. To increase the genetic basis is necessary to incorporate new sources of variability such as lines from a germplasm bank, one of the strategies for incorporating exotic lines is the phenotypic characterization of these genotypes.This work was the phenotypic characterization of 80 exotic soybeans lines and 15 Brazilian commercial checks. Ten agro morphological traits were evaluated in four trials in the 2012/2013 and 2013/2014 seasons. Deviance analyzes were performed, Principal Component and two methods of clustering: Ward's and Average Linkage. All variables showed significance for the deviance analysis in the analyzed genotypes, indicating good genetic variability between them. Much of the total variance presented was explained by three main principal components, however the first two components allowed greater power of differentiation among the genotypes. In Ward's cluster were formed ten groups, and in the Average Linkage method the lines were divided into seven groups. Based on the results it was concluded that, lines of this soybean panel have genetic divergence, based on agro morphological markers, so the

selection of genotypes for use in soybean breeding programs aimed at increasing genetics variability is possible

Keywords: *Glycine max*; Clustering; Exotic germplasm

3.1 Introdução

Existe diversidade genética suficiente para sustentar os ganhos de produtividade no melhoramento de soja? Este tipo de questão permanece sem resposta tanto para a soja quanto para as outras culturas. Os melhoristas tentam minimizar custos e tempo cruzando materiais elite, entretanto este tipo de abordagem tende a ocasionar o estreitamento da base genética das espécies cultivadas causando consequências graves como patamares de produtividade e vulnerabilidade genética (Committee on Genetic Vulnerability, 1972; RODGERS et al., 1983; SMITH, 1988; SMITH et al., 1992).

Segundo dados da CONAB (CONAB, 2015) durante os últimos 35 anos, a área cultivada com soja teve um crescimento de 248% e um aumento de produtividade de 506%, valor este cerca de duas vezes maior que o primeiro. Sabe-se que um dos fatores mais importantes atribuídos para a evolução da produtividade ao longo deste período foi o melhoramento genético (SILVA NETO, 2011.). Entretanto, este incremento no rendimento de grãos pode estagnar nos próximos anos, principalmente se os melhoristas continuarem a utilizar os mesmos parentais em seus cruzamentos. Vários estudos comprovam o estreitamento da base genética da cultura da soja no país (HIROMOTO e VELLO, 1986; PRIOLLI et al., 2004; WYSMIERSKI e VELLO, 2013).

A fim de aumentar a base genética desta cultura, deve-se conhecer as características dos materiais disponíveis dentro dos bancos de germoplasma, visando a incorporação destes em programas de melhoramento. Apesar do desenvolvimento de técnicas moleculares para acessar a diversidade presente entre os acessos, o uso de dados fenotípicos pode estar mais associado com o valor genético (*Breeding Value*) do que os dados de marcadores, pois muitos marcadores estão em regiões não codantes do DNA, já os genes controlando as características expressadas fenotipicamente sim (CARTER et al., 2004). Sendo assim, é de fundamental importância a caracterização fenotípica da diversidade genética.

Existem várias metodologias para o estudo de diversidade genética fenotípica, podendo receber destaque as técnicas de análise multivariada, as quais permitem a análise de várias características simultaneamente. Dentre as abordagens mais utilizadas podemos citar as técnicas de agrupamento, análise de componentes principais, análise discriminante e fatorial.

Na cultura da soja a diversidade fenotípica tem sido estuda há muito tempo, principalmente antes do desenvolvimento das ferramentas moleculares. Na literatura podemos encontrar vários exemplos envolvendo este tipo de trabalho. Perry e Macintosh (1991) avaliaram a diversidade genética de 2.250 acessos provenientes de 78 países ao redor do globo utilizando 17 características morfológicas para avaliar a diversidade entre acessos e regiões geográficas. Foram empregados neste trabalho, análise discriminante e o índice de diversidade de Shanon-Weaver e com isso, os autores observaram variação fenotípica para quase todos os caracteres nas diferentes regiões analisadas. Já Sneller (1994) utilizou coeficiente de parentesco, análise de componentes principais e agrupamento para analisar 122 linhagens elites dos Estados Unidos com o objetivo de identificar padrões de diversidade dentro do país. O autor detectou pequenas variações entre as linhagens de programas de melhoramento público e privado, além de poucas alterações de diversidade genética em relação a estudos anteriores.

No Brasil alguns estudos com diversidade fenotípica em soja também foram feitos. Villela (2013) estudou a diversidade presente em 74 cultivares de soja oriundas de diferentes programas de melhoramento, através de dez caracteres agro morfológicos avaliadas em análises de agrupamento e componentes principais, sete grupos distintos foram detectados. Ferreira Junior et al. (2015) avaliou a diversidade e o desempenho de linhagens de soja oriundas de cruzamentos, através de 11 características fenotípicas. O autor encontrou seis grupos pelo método de agrupamento de Ward, indicando assim a presença de variabilidade genética.

Em 2010, Mulato e colaboradores na Escola Superior de Agricultura “Luiz de Queiroz”, iniciaram um trabalho para a caracterização genética e genotípica da diversidade genética de um painel de acessos de soja contendo 79 linhagens exóticas de soja representando diferentes partes do globo. Estes mesmos acessos juntamente

com mais 16 testemunhas comerciais foram analisados neste trabalho, o qual tem como objetivo avaliar a diversidade fenotípica entre estes acessos utilizando características fenotípicas.

3.2 Material e Métodos

Um painel de acessos composto por 95 genótipos de soja, entre estes, testemunha comerciais e PI's (*Plant Introductions*) foram analisados neste trabalho (ANEXO A). O primeiro experimento foi conduzido na safra 2012/2013 em Piracicaba-São Paulo (SP), Jaboticabal-SP, e Ponta Grossa-Paraná (PR). E o segundo experimento foi instalado nas cidades de Piracicaba-SP e Jaboticabal-SP na safra 2013/2014.

O delineamento experimental utilizado foi Alfa Látice 5x19, com 3 repetições e parcelas de 4 linhas de 5 metros com espaçamento entre estas de 0,5 m. Apenas as duas linhas centrais de cada parcela foram colhidas evitando assim efeitos de bordadura e mistura varietal entre os acessos.

Os seguintes caracteres agro morfológicos foram avaliados:

- Altura da planta no florescimento (APF) – Média das cinco plantas centrais da parcela, medidas do solo ao final da haste em cm.
- Período de granação (PEG) – Número de dias entre os estágios R5 (número de dias da semeadura até o início de enchimento de grãos) e R7 (número de dias da semeadura até a granação completa).
- Número de dias para a maturidade (NDM) – Número de dias da semeadura até 95% das vagens maduras em 50% da parcela.
- Altura da planta na maturidade (APM) – Média das cinco plantas centrais da parcela, medidas do solo ao final da haste em cm.
- Valor agronômico (VA) – Escala de notas de 1 a 5, sendo 1 para planta com baixo valor e 5 para planta excelente, para a arquitetura geral das plantas na parcela.
- Acamamento (AC) – Escala de notas de 1 a 5, sendo 5 para planta totalmente acamada e 1 para planta ereta.

- Altura da inserção da primeira vagem (IPV) – Média das cinco plantas centrais da parcela, medidas do solo até a primeira vagem da haste em cm.
- Produtividade de grãos (PROD) – Massa total das sementes produzidas na parcela, em quilograma por hectare (kg ha^{-1}).
- Massa de cem sementes (MCS) – Massa de 100 sementes em gramas (g).

Teor de óleo (OLEO) – média em porcentagem de três leituras no espectômetro NIR (*Near-infra red spectroscopy*

Os dados de todas as características avaliadas, foram analisados pelo software Selegen-Reml/Blup (Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos), desenvolvido por Resende e colaboradores em 1994. Foi utilizado um modelo misto em blocos incompletos em vários locais e uma só colheita. A significância dos efeitos do modelo foi testado pela Análise de Deviance (ANADEV), e a significância dos efeitos testadas pelo Qui-Quadrado a 1% e 5% (ANEXO B).

As médias obtidas pela ANADEV nos 5 locais avaliados para cada característica, foram utilizadas para realização das análise multivariadas. Estas médias foram padronizadas minimizando os efeitos de *outliers* na análise conforme procedimento descrito por Huber (1973).

Três metodologias multivariadas foram empregadas, a Análise de Componentes Principais (ACP), e dois tipos de Agrupamento hierárquico, *Ward* e *Average Linkage*. O princípio da ACP é sumarizar a variação total explicada pelas variáveis originais em um novo conjunto de variáveis transformadas (componentes principais) com o mesmo número de dimensões de tal maneira que as primeiras componentes principais sempre explicarão maior variação que as componentes principais subsequentes. Através da projeção das amostras “ou acessos” em relação aos componentes principais é possível realizar um agrupamento em que é admissível analisar a relação entre as amostras estudadas. Já a análise de agrupamento, ou *clustering*, tem como objetivo o agrupamento de objetos ou indivíduos com características semelhantes entre si. Esta análise é feita a partir de uma matriz de dados de similaridades ou distâncias (dissimilaridades), tais medidas por sua vez são calculadas considerando os valores das amostras para cada variável.

Para o cálculo dos componentes principais, foi utilizada a matriz de correlações, já que as variáveis analisadas possuem escalas diferentes e com a utilização de correlações não há necessidade de transformação dos dados.

O cálculo de correlação cofenético (BARROSO e ARTES, 2003; CRUZ e CARNEIRO, 2003)para os métodos de agrupamento, e a matriz de distâncias Euclidianas utilizada em ambos os métodos de agrupamento, foi calculada a partir das médias ajustadas dos 5 ambientes para as 10 características avaliadas no software Genes (CRUZ, 2001).

O método de agrupamento de *Ward's* utiliza como distância entre os clusters a seguinte fórmula (MILLIGAN, 1980):

$$D_{KL} = \frac{\|\overline{x_K} - \overline{x_L}\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

Sendo que:

C_K é o k ésimo cluster;

C_L é o L ésimo cluster;

$\overline{x_K}$ é o vetor de médias para o cluster C_K;

$\overline{x_L}$ é o vetor de médias para o cluster C_L;

N_k é o número de observações em C_K;

N_L é o número de observações em C_L;

O método de agrupamento de *Average Linkage,* conhecido como UPGMA*,* utiliza como distância entre os clusters a seguinte fórmula (SOKAL e MICHENER, 1958):

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d\ (x_i, x_j)}{N_K N_L}$$

Sendo que:

x_i é a i ésima observação;

x_j é a j ésima observação;

C_K é o k ésimo cluster;

C_L é o L ésimo cluster;

N_k é o número de observações em C_K;

N_L é o número de observações em C_L;

$\|x\|$ é a raiz quadrada da soma de quadrados dos elementos de x(O comprimento da distância Euclidiana do vetor x);

$d\ (x_i, x_j)$ é igual a $\|x_i - x_j\|^2$

Em ambos os métodos de agrupamento o número de clusters foi definido pela estatística *Cluster Cubic Criterion* (CCC) descrito por Sarle (1983). Esta estatística basicamente assume que todos os clusters foram obtidos de uma distribuição uniforme, assim, valores esperados de R^2 baseados nesta distribuição são então comparados com valores observados de R^2.

As análises de componentes principais e de agrupamento foram feitas pelo software JMP versão 12 (SAS INSTITUTE, 2015).

3.3 Resultados e Discussão

A análise de componentes principais foi feita a fim de verificar o agrupamento dos indivíduos de acordo com as características avaliadas. Observa-se que os três primeiros componentes principais explicaram 71,53% da variação fenotípica total apresentada entre os acessos (Tabela 1). Segundo o critério de Kaiser (1958) deve-se reter apenas aqueles componentes os quais apresentem autovalores maiores que 1, ou seja, no caso deste estudo os componentes 1, 2 e 3 (CP1, CP2 e CP3 respectivamente).

Os auto vetores (Tabela 2) explicam o quanto cada variável original está associada a cada um dos componentes principais. No CP1 os caracteres altura de planta na maturação (APM) e altura da planta no florescimento (APF), foram as que mais contribuíram para a formação deste primeiro componente. Já no CP2, destacam-se as variáveis número de dias para a maturação (NDM) e produtividade de grãos (PROD). Por sua vez na formação do terceiro componente as características inserção da primeira vagem (IPV) e teor de óleo (OLEO), contribuíram com os maiores valores.

Tabela 1. Autovalores, variância e variância acumulada em relação a total, na análise de componentes principais em 95 linhagens de soja

Componentes	Auto valores	Variância	Variância Acumulada(%)
1	4,18	41,84	41,84
2	1,96	19,64	61,48
3	1,01	10,06	71,53
4	0,79	7,90	79,43
5	0,77	7,69	87,11
6	0,45	4,53	91,64
7	0,38	3,83	95,47
8	0,20	1,97	97,45
9	0,17	1,72	99,16
10	0,08	0,84	100,00

As características que mais contribuíram do CP1(APF, AC, APM e IPV), estão relacionadas a arquitetura da planta. Para CP2, todas as características, exceto o caráter acamamento (AC), sendo um bom indicativo para a utilização do segundo componente como diferenciador dos genótipos. O CP3, por sua vez, foi influenciado por poucas variáveis, sendo por este motivo não levado em consideração na análise gráfica.

Este mesmo tipo de relação entre as variáveis pode ser observado no círculo de correlações apresentado na Figura 1b. Este gráfico, permite a observação da correlação entre as variáveis e sua importância nos dois primeiros componentes principais, quanto mais próximas aos eixos x ou y do círculo mais as características estão associadas com os componentes principais plotados respectivamente. Observa-se pelo gráfico que as variáveis AC e MCS estão altamente relacionadas ao componente 1, apesar de terem sentidos opostos devido aos sinais, sendo AC com contribuição alta e positiva (0.4352) e MCS com contribuição alta e negativa (0.4352).

As correlações entre as características também podem ser observadas pela proximidade destas no plano, como por exemplo, as variáveis PROD e PEG. Ainda na figura 1b, podem-se observar três grupos de variáveis, o primeiro grupo englobando as variáveis NDM, APF, APM, IPV, o segundo grupo os caracteres PROD, PEG, VA, OLEO e MCS e por fim o terceiro representado pelo acamamento de plantas.

Tabela 2. Autovetores da análise de componentes principais para 10 características avaliadas fenotipicamente em 95 genótipos de soja

	Componentes							
Caráter	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**	**CP8**
APM	0,38	0,3354	-0,062	0,133	0,138	0,3897	0,2366	0,2938
PEG	-0,261	0,396	-0,066	-0,455	-0,463	0,0923	0,2542	0,2457
VA	-0,358	0,3188	-0,081	0,236	0,3518	-0,11	-0,198	0,6574
AC	0,4352	-0,058	-0,064	0,1376	-0,089	0,1606	0,5453	0,1408
MCS	-0,381	0,0382	0,2235	-0,116	0,1585	0,7964	-0,043	-0,242
NDM	0,2474	0,4932	-0,045	-0,182	-0,387	-0,075	-0,398	-0,15
IPV	0,1685	0,1764	0,7334	-0,407	0,3721	-0,266	0,1591	0,0262
APF	0,3728	0,3241	0,0377	0,2267	0,1904	0,1679	-0,407	-0,188
OLEO	-0,2	0,1552	0,5447	0,6484	-0,437	-0,065	0,1259	-0,055
PROD	-0,238	0,4699	-0,305	0,1283	0,3082	-0,243	0,4192	-0,53

APM – altura da planta na maturação, PEG – período de granação, VA – valor agronômico, AC – acamamento, MCS – massa de 100 sementes, NDM – número de dias para a maturação, IPV – inserção da primeira vagem, APF – altura da planta no florescimento, OLEO – Teor de óleo, PROD – produção de grãos

A Figura 1a representa o comportamento dos 95 genótipos de soja em relação aos dois primeiros componentes principais. As testemunhas comerciais utilizadas, V-máx, Sambaíba, Potência entre outras, ficaram agrupadas em um só grupo localizadas no sentido de crescimento da variável produtividade de grãos. As PI's, PI 145079, PI 170889 e PI 281898 também foram alocadas neste grupo podendo ser um indicativo de boa produtividade de grãos e demais características correlacionadas nestes acessos.

Um segundo grupo formado por outras PI's dentre elas PI 159922, PI 417563 e PI 259540 estão positivamente influenciadas pelas variáveis de arquitetura APF e APM, e NDM, sendo a correlação alta entre estas variáveis explicada por Lopes et al. (2012), segundo os autores quanto maior o ciclo, maior a planta e os internódios produzidos, por isso as correlações altas entre estas variáveis. O grupo no quarto quadrante está altamente associado ao aumento da característica acamamento, sendo esta uma característica negativa, pois as notas de AC mais altas indicam plantas acamadas e tendem possuir baixo valor agronômico, por dificultarem a colheita mecânica devido ao porte não ereto e grãos danificados devido ao contato com o solo.

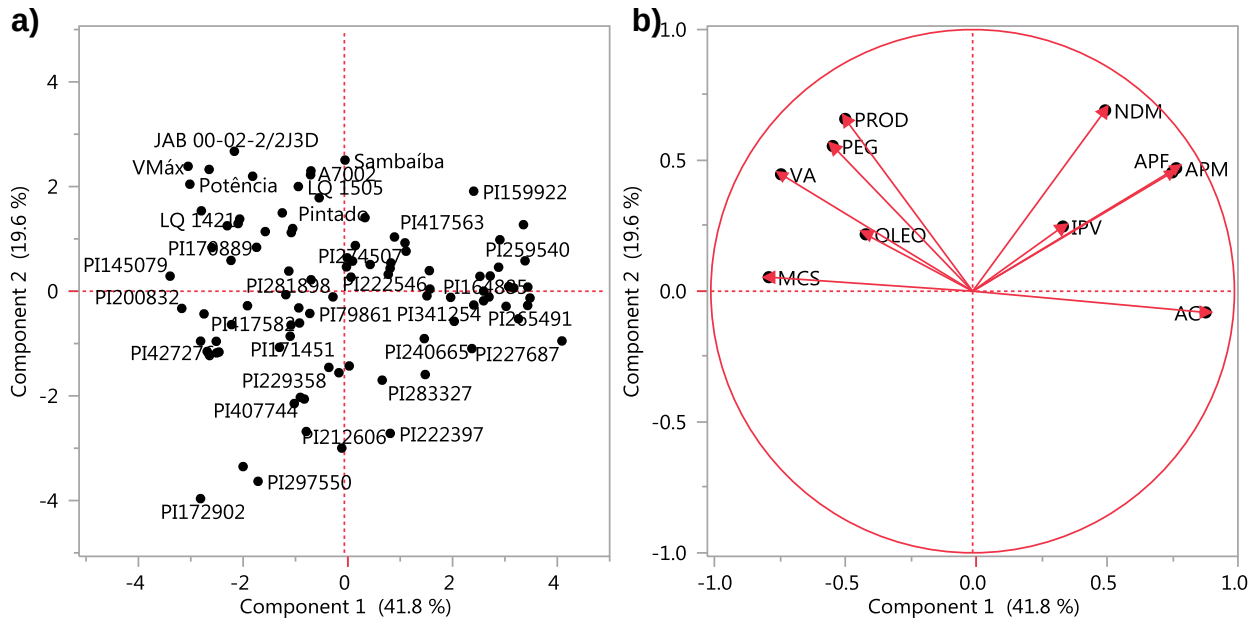

Figura 1 – a) Gráfico dos dois primeiros componentes principais para as 95 cultivares de soja; b) Círculo de correlações para as 10 características agro morfológicas analisadas

Mesma abordagem utilizada neste trabalho foi feita por Marconato (2014), utilizando 93 genótipos dos 95 deste trabalho. Realizando análise de componentes principais em nove características agro morfológicas foram observados resultados similares ao deste estudo. Sendo os três primeiros componentes responsáveis por 71% da variância fenotípica total. A autora encontrou também grande associação entre o componente três e a característica inserção da primeira vagem, devido a isto este componente não possibilitou a discriminação entre os genótipos devido a baixa correlação entre IPV e PROD, fato corroborado por outros estudos que obtiveram valores baixos para este mesmo tipo de correlação (MUNIZ et al., 2002; ALCANTARA NETO et al., 2011).

Além da análise de componentes principais outro método multivariado empregado foi a análise de agrupamento hierárquico, utilizando duas metodologias: *Average Linkage* ou UPGMA (SOKAL e MICHENER, 1958), e o método de *Ward's* (MILLIGAN, 1980). Ambos os métodos possuem vantagens e desvantagens, o primeiro é sensível em relação a dados com mesma variância, e o segundo viesado em relação à produção de clusters com o mesmo número de observações.

O dendrograma do agrupamento de Ward's está apresentado na Figura 2. Dez *clusters* foram formados com base nas dez características agro morfológicas utilizadas.

As cultivares brasileiras foram alocadas nos *clusters* 2 e 3. No cluster dois além das LQ´s do programa de melhoramento da ESALQ, foram agrupadas as cultivares IAC100 e Paranagoinia também do Brasil. Além destas linhagens, quatro acessos exóticos foram alocados neste cluster, a PI 210352 originária de Moçambique com valores médios de NDM 131 dias, VA 3,17, AC 1,17 e PROD de 2088 kg ha^{-1}, a PI 281898 originária da Malásia com valores médios de NDM 135 dias, VA 2,92, AC 1,75 e PROD 1296, e os acessos Kinoshita e Shiranui mencionados na literatura como fontes de resistência a ferrugem asiática da soja (*Phakopsora pachyrhizi*) (BATISTA, 2008).

No *cluster* 3 (Figura 2) ficaram alocadas as demais cultivares brasileiras, entre estas linhagens, os valores médios para produtividade de grãos ultrapassaram os 2000 kg ha^{-1}, variando de 2141 kg ha^{-1} para a cultivar A7002 à 3332 kg ha^{-1} de Potência. Os valores apresentados para OLEO ultrapassaram os 18%, variando de 18,85% para a cultivar A7002 e 21,21% para a linhagem JAB 00-02-2/2J3D. Neste cluster constam as linhagens com maior média de produtividade e valor agronômico já que em sua maioria são linhagens adaptadas as regiões analisadas nos experimentos.

Algumas características podem ser destacadas nos *clusters* formados. O *cluster* seis apresentou acessos exóticos com valores médios baixos para notas de AC, variando de 1,33 à 2,78. Assim, as linhagens alocadas neste grupo apresentaram porte mais ereto que as demais. Em contrapartida as linhagens do cluster nove apresentam valores altos para notas de AC variando de 3,13 à 4,11. É interessante ressaltar que os acessos do *cluster* seis apresentaram também menor APM, 44,18 cm à 74,49 cm ao oposto que, no *cluster* 9 os valores médios de APM variaram de 11,03 cm à 144,18 cm. Buzello et al (2013), estudando plantas de soja encontrou correlação positiva entre as características acamento e altura de plantas, refletindo que uma redução na altura de plantas consequentemente promove uma redução no acamamento.

Além dos valores de AC, os acessos exóticos do *cluster* 9 também possuem em comum valores médios de OLEO menores que 20%, variando de 14,94% na PI PI222550 à 18,05% na PI204333.

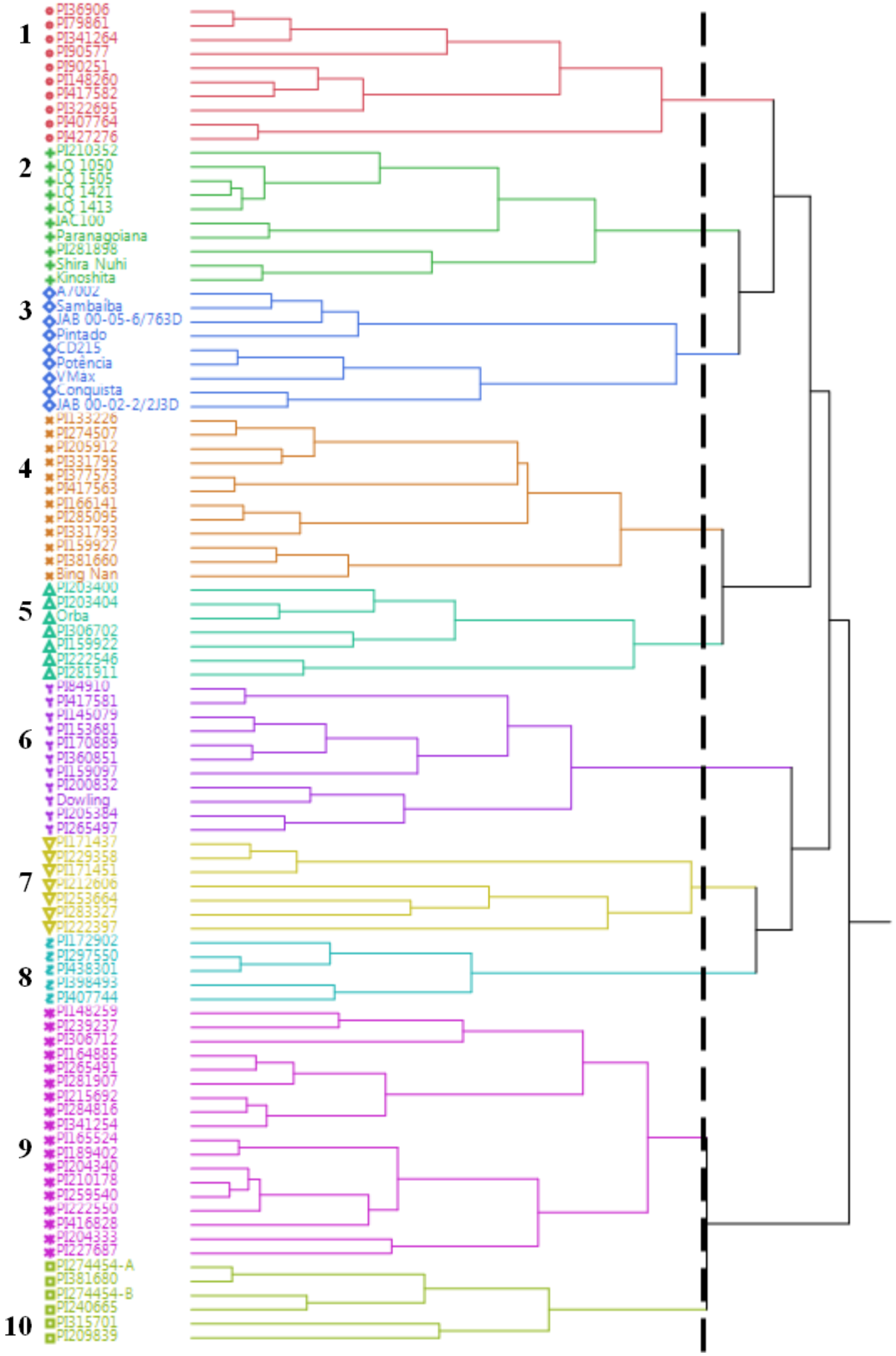

Figura 2 – Dendrograma gerado pelo método de agrupamento de Ward em 95 linhagens de soja. O número de clusters = 10 obtido pelo método CCC representado pela linha tracejada

As 95 linhagens analisadas (ANEXO A) no agrupamento são originárias de 37 países ao redor do mundo. Entretanto os *clusters* formados não correspondem as mesma regiões de origem dos acessos. Os acessos provenientes da China e Japão foram agrupados em quase todos os *clusters*, podendo ser uma explicação válida, que a região compreendida por estes países é considerada centro de diversidade e domesticação da soja (OLIVEIRA et al., 2010; LI; NELSON, 2001).

Pelo método *Average Linkage*, foram formados 7 *clusters* utilizando o critério de CCC. Os *clusters* estão representados na forma de dendrograma na Figura 3. O número de *clusters* formados foi diferente do método de *Ward's*. As cultivares brasileiras foram agrupadas em dois *clusters* diferentes*,* assim como no método de *Ward's* (Figura 3). Neste tipo de agrupamento as cultivares Potência, Vmáx, Conquista, CD 215 e JAB 00-02-2/2J3D foram alocadas no *cluster* 2, enquanto que as demais linhagens do Brasil foram alocadas no *cluster 1*.

O *cluster* 2 agrupou linhagens brasileiras com médias variando de 2815 kg ha^{-1} para a cultivar Conquista e 3332 kg ha^{-1} para a cultivar Potência. Os valores AC variaram de 1,22 para VMáx à 1,62 em Conquista, as notas para VA variaram de 3,50 à 3,89 e o NDM variou de 132 à 144 dias.

O maior agrupamento formado, o *cluster* 1, alocou dez cultivares brasileiras. Os valores médios neste cluster para a característica PROD, variaram de 1590 kg ha^{-1} para a PI341264 à 2905 kg ha^{-1} para a cultivar Paranagoiania, para AC as notas variaram de 1,28 na PI210352 à 3,64 para a PI306702, já para a característica VA os valores variaram de 1,89 para a PI306702 à 3,57 na linhagem LQ 1403. As melhores notas de VA e AC foram observadas nas cultivares brasileiras.

Alguns clusters formados no agrupamento de *Ward's* foram semelhantes aqueles do método *Average Linkage*. No *cluster* 3 (Figura 3) foram alocadas 13 linhagens, pelo método de Average Linkage, dentre estas, 11 genótipos são comuns ao *cluster* seis do agrupamento de Ward's. Os dois genótipos não inclusos neste último cluster foram os acessos exóticos PI407764 e PI427276 originárias da China. Outros clusters semelhantes, com pequenas diferenças de acessos entre os agrupamentos, foram os *cluster* 9 (Figura 2) e *cluster* 5 (Figura 3).

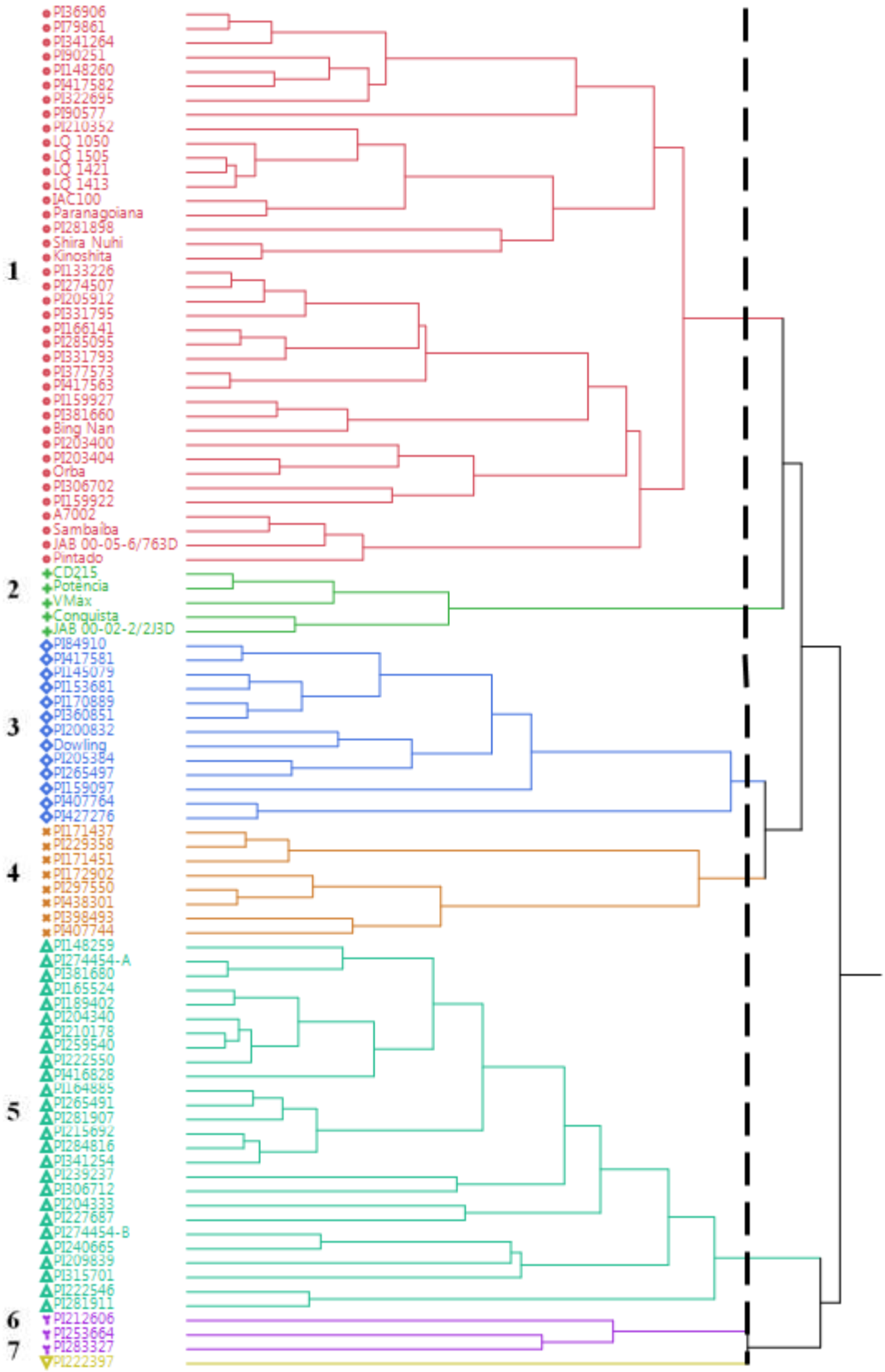

Figura 3. Árvore pelo método de agrupamento *Average Linkage* dos 95 linhagens de soja. O número de clusters = 10 obtido pelo método CCC representado pela linha tracejada

A PI222397 originária do Paquistão foi o único acesso a ser alocada em um *cluster* sozinha. Esta linhagem apresentou valores para as características agro morfológicas como PROD 916 kg ha^{-1}, AC 4,22, VA 1,61 e APM 60,92. Estas características permitiram a diferenciação desta linhagem dos demais acessos no agrupamento de *Average Linkage*.

O *cluster* 4 (Figura 3) apresentou as cultivares com menores valores de AC, variando de 1,17 na PI438301 à 1,88 na PI407744. Os valores de APM também apresentaram valores baixos confirmando como mencionado anteriormente a correlação positiva entre AC e APM. O oposto ocorre no *cluster* 5, onde os valores de AC variaram de 2,56 na PI133226 à 4,44 para a PI227687. Além disso os valores médios para a característica APM foram altos, variando de 88 cm à 144 cm.

Mulato (2010), utilizando os mesmos acessos exóticos deste trabalho, avaliou a diversidade genética via marcadores agro morfológicos. O autor avaliou quatorze características quantitativas e 6 variáveis categoricas ou qualitativas, utilizando o agrupamento de Tocher e variáveis canônicas. Na análise dos caracteres qualitativos pelo agrupamento de Tocher foram formados dezesseis grupos, já quando analisados os caracteres quantitativos foram formados vinte grupos. Na variável canônica explicando a maior porcentagem da variabilidade total (76,98%) estão os caracteres NDM e início da granação (R5).

Outros trabalhos também já foram feitos visando identificar a diversidade genética em soja a partir de caracteres agro morfológicos. Cui et al. (2001), observaram a separação em grupos através do método *Average Linkage* em cultivares de soja americanas e chinesas. Similarmente utilizando a mesma metodologia Nogueira (2011) conseguiu diferenciar 90 linhagens de soja de acordo com as regiões de adaptação destas.

Visando detectar as diferenças entre os dois tipos de agrupamento, utilizou-se o cálculo da correlação cofenética. Este tipo de correlação mede o ajuste entre a matriz de similaridade original e matriz ajustada pelo método de agrupamento. O coeficiente varia entre 0 e 1 e quanto mais próximo de 1 menor a distorção do método de agrupamento nos dados originais (CRUZ e CARNEIRO, 2006). O coeficiente de correlação para o método de *Ward`s* foi de 0,61 e já para o método Average Linkage foi

0,67, assim o último tipo de agrupamento representou um melhor ajuste dos dados implicando em uma menor distorção no dendrograma apresentado.

Os dois métodos de agrupamentos hierárquicos possuem similaridades no agrupamento de vários acessos como discutido nos resultados anteriormente. Kantety et al (1995), utilizou cinco métodos de agrupamento, *Single Linkage, UPGMA, UPGMC, Complete Linkage* e *Ward's* para avaliar similaridade entre linhagens de milho pipoca utilizando marcadores microsatélites. Os diferentes métodos apresentaram resultados similares com mínimas diferenças entres eles, entretanto segundo os autores o método que se mostrou mais consistente com os grupos heteróticos já conhecidos foi o método de UPGMA.

3.4 Conclusões

Os acessos deste painel de soja possuem divergência genética, com base em marcadores agro morfológicos, sendo assim, vemos que é possível a seleção de genótipos para utilização em programas de melhoramento de soja visando o aumento da variabilidade genética.

Os dois métodos de agrupamento utilizados apresentaram diferenças no número de clusters. O método Average Linkage apresentou uma maior valor de correlação cofenética e portanto representou um melhor ajuste dos dados originais

É recomendado, que se complete este estudo com a caracterização da diversidade genética utilizando marcadores moleculares, a fim de comprovar os resultados obtidos no presente trabalho e fornecer uma informação completa sobre a variabilidade genética presente neste painel de acessos de soja.

Referências

ALCANTARA NETO, F. DE; GRAVINA, G. DE. A.; MONTEIRO, M. M. DE. S.; ORAIS, F. B. DE.; PETTER, F. A.; ALBUQUERQUE, J. A. A. DE. Análise de trilha do rendimento de grãos de soja na microrregião do Alto Médio Gurguéia. **Comunicata Scientiae**, Bom Jesus, v. 2, p.107-112, 2011.

BARROSO, L.P.; ARTES, R. **Análise multivariada**. Lavras: UFLA, 2003. 151p.

BATISTA, C. E. de A. **Mapeamento de genes associados à resistência da soja a ferrugem asiática (*Phakopsora pachyrhizi*)**. 2008. 57 f. Dissertação (Mestrado) - Curso de Genética e Melhoramento de Plantas, Universidade de São Paulo - "escola Superior de Agricultura Luiz de Queiroz", Piracicaba, 2008.

BUZZELLO, G. L.; TREZZI, M. M.; MARCHESE, J. A.; XAVIER, E.; MIOTTO JUNIOR, E.; PATEL, F.; DEBASTIANI, F. Action of auxin inhibitors on growth and grain yield of soybean. **Revista Ceres**, Viçosa, v. 60, n. 5, p.621-628, 2013.

CARTER, T. E.; NELSON, R. L.; SNELLER, C. H.; CUI, Z. Genetic Diversity in Soybean. In: BOERMA, H. R. (Ed.); SPECHT, J. E. Soybean: **improvement, production and uses**. 3. ed. Madison: American Society Of Agronomy, 2004. Cap. 8. p. 303-396.

COMMITTEE ON GENETIC VULNERABILITY OF MAJOR CROPS. **Genetic vulnerability of major crops**. Washington: National Academy Of Sciences, 1972. 307 p.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. Séries históricas safras 1976/77 a 2014/2015. Disponível em: <http://www.conab.gov.br/conteudos.php?a=1252&t=&Pagina_objcmsconteudos=3#A_objcmsconteudos>. Acesso em: 28 ago. 2015.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 2003. 585p.

CRUZ, C. D.. **Programa genes: aplicativo computacional em genética e estatística.** Viçosa: UFV, 2001. 648 p.

CUI, Z.; CARTER, J. R.; BURTON, J. W.; WELLS, R. Phenotypic diversity of modern Chinese and North American soybean cultivars. **Crop Science Society Of America**, Madison, v. 41, n. 6, p.1954-1967, 2001.

FERREIRA JÚNIOR, J. A., UNÊDA-TREVISOLI, S. H., ESPÍNDOLA, S. M. C. G., VIANNA, V. F., MAURO, A. O. DI. (2015). Diversidade genética em linhagens avançadas de soja oriundas de cruzamentos biparentais, quádruplos e óctuplos1. **Revista Ciência Agronômica**, Fortaleza, v. 46, p.339-351, 2015.

HIROMOTO, D. M.; VELLO, N. A.. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazilian Journal Of Genetics,** Ribeirão Preto, v. 09, n. 2, p.295-306, 1986.

HUBER, P. J. Robust Regression: Asymptotics, Conjecture, and Monte Carlo. **Annals Of Statistics**, Hayward, v. 1, n. 5, p.799-821,1973.

KANTETY, R. V.; ZENG, X.; BENNETZEN, J. L.; ZEHR, B. E. Assessment of genetic diversity in dent and popcorn (Zea mays L.) inbred lines using inter-simple sequence repeat (ISSR) amplification. **Molecular Breeding**, Dordrecht, v. 1, n. 4, p.365-373, 1995.

LI, Z.; NELSON, R. L. Genetic Diversity among Soybean Accessions from Three Countries Measured by RAPDs. **Crop Science Society Of America**, Madison, v. 41, n. 4, p.1337-1347, 2001.

MARCONATO, M. B. **Diversidade fenotípica por meio de caracteres agronômicos em acessos de soja**. 2014. 50 p. Dissertação (mestrado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, 2014.

MILLIGAN, G. W.. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. **Psychometrika**, Williamsburg, v. 45, p.325-342, 1980.

MULATO, B. M.; MÖLLER, M.; ZUCCHI, M. I.; QUECINI, V.; PINHEIRO, J. B. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. **Pesquisa Agropecuária Brasileira**., v. 45, n. 3, p. 276-283, 2010.

MUNIZ, F. R. S.; DI MAURO, A. O.; UNÊDA-TREVISOLI, S. H.; OLIVEIRA, J. A.;BÁRBARO, I. M.; ARRIEL, N. H. C.; COSTA, M. M. Parâmetros genéticos e fenotípicos em populações segregantes de soja. **Revista Brasileira de Oleaginosas e Fibrosas**, Campina Grande, v.6, n.3, p.615-622, 2002.

OLIVEIRA, M. F.; NELSON, R. L.; GERALDI, I. O.; CRUZ C. D.; DE TOLEDO, J. F. Establishing a soybean germplasm core collection. **Field Crops Research**, Amsterdam, v. 119, n. 2-3, p.277-289, 2010.

PERRY, M. C.; McINTOSH, M. S. Geographical patterns of variation in the USDA soybean germplasm collection: I. Morphological traits. **Crop Science Society Of America**, Madison, v. 31, n. 5, p. 1350-1355, 1991.

PRIOLLI, R. H. G.; MENDES-JUNIOR, C. T.; SOUSA, S. M. B.; ARANTES, N. E.; CONTEL, E. P. B. Diversidade genética da soja entre períodos e entre programas de melhoramento no Brasil. **Pesquisa Agropecuária Brasileira,** Brasília, v. 39, n. 10, p.967-975, 2004.

RODGERS, D. M.; MURPHY, J. P.; FREY, K. J. Impact of Plant Breeding on the Grain Yield and Genetic Diversity of Spring Oats. **Crop Science Society Of America,** Madison, v. 23, n. 4, p.737-740, 1983.

SARLE, W. S. **Cubic Clustering Criterion**. Cary: Sas Institute Inc, 1983. 59 p. (SAS Technical Report A-108).

SILVA NETO, S. P. DA. **A evolução da produtividade da soja no Brasil**. 2011. EMBRAPA CERRADOS. Disponível em: <http://www.cpac.embrapa.br/noticias/artigosmidia/publicados/335/>. Acesso em: 24 ago. 2015.

SMITH, J. S. C. Diversity of United States Hybrid Maize Germplasm; Isozymic and Chromatographic Evidence. **Crop Science Society Of America**, Madison, v. 28, n. 1, p.63-69, 1988.

SMITH, J. S. C.; SMITH, O. S.; WRIGHT, S.; WALL, M. W. Diversity of U.S. Hybrid Maize Germplasm as Revealed by Restriction Fragment Length Polymorphisms. **Crop Science Society Of America**, Madison, v. 32, n. 3, p.598-604, 1992.

SNELLER, C. H.. Impact of transgenic genotypes and subdivision on diversity within elite North American soybean germplasm. **Crop Science Society Of America**, Madison, v. 38, p.409-414, 2003.

SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. **University Of Kansas Scientific Bulletin**, Lawrence, v. 38, p.1409-1438, 1958.

VILLELA, O. T. **Diversidade fenotípica e molecular de cultivares brasileiras de soja portadoras de gene RR**. 2013. 67p. Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, 2013.

WYSMIERSKI, P. T.; VELLO, N. A. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. **Genetics And Molecular Biology**, Ribeirão Preto, v. 36, n. 4, p.547-555, 2013.

4 ESTRUTURA DE POPULAÇÃO E DIVERSIDADE GENÉTICA DE ACESSOS DE SOJA UTILIZANDO GENOME-WIDE SNPs.

Resumo

Uma boa caracterização de acessos exóticos pertencentes a bancos de germoplasma pode ajudar na melhor utilização dos recursos genéticos e incorporação destes genótipos como fonte de variabilidade em programas de melhoramento. Um painel de soja contendo 80 acessos exóticos e 15 linhagens brasileiras foram caracterizadas molecularmente utilizando 10017 SNPs do chip de genotipagem *Axiom® Soybean Genotyping Array*. Foi detectada a presença de diversidade genética entre os acessos, com uma diversidade gênica variando de 0 a 0,34, um PIC variando de 0,021 a 0,480 e Ho de 0 a 0,34. Na análise de agrupamento pelo método *Neighbor Joining Tree* foi possível detectar e presença de dois grupos utilizando 1000 *bootstraps* para produção de um dendrograma consenso. A estrutura da população foi analisada por duas abordagens Bayesiana (STRUCTURE) e Análise Discriminante de Componentes Principais. Os resultados mostram um número ótimo de k=2 para as duas abordagens,contando com pequenas diferenças nas alocações dos indivíduos.

Palavras chave: *Glycine max*; Germoplasma; Marcadores moleculares; DAPC

Abstract

A good characterization of exotic lines of the germplasm banks can help make better use of genetic resources and incorporation of these genotypes as a source of variability in breeding programs. A soybean panel containing 80 exotic lines and 15 Brazilian improved lines were characterized molecular using 10017 SNPs of the genotyping chip Axiom® Soybean Genotyping Array. The presence of genetic diversity among the accessions was detected with a gene diversity ranging from 0 to 0.34, a PIC ranging from 0.021 to 0.480 and a Ho 0 to 0.34. In the clustering analysis using the Neighbor Joining Tree method was detected the presence of two clusters using 1000 bootstraps to produce a consensus dendrogram. The population structure was analyzed by two approaches, Bayesian (STRUCTURE) and Discriminate Analysis of Principal Component. The results show a great number of k = 2 for the two approaches, with minor differences in allocations of individuals.

Keywords: *Glycine max*; Germplasm; Molecular markers; DAPC

4.1 Introdução

É de fundamental importância o entendimento sobre a diversidade genética presente nos bancos de germoplasma e o parentesco entre as cultivares comerciais modernas e seus parentes exóticos. Tais cultivares crioulas são os depositórios

naturais da variabilidade genética presente nas espécies cultivadas, contendo fontes de resistência a fatores abióticos como tolerância a seca, calor e frio e fatores bióticos, como doenças e pragas, sendo, portanto, fontes para aumento da diversidade genética em programas de melhoramento. Entretanto, estima-se que em soja, 0,6% dos acessos totais presentes nos bancos de germoplasma no mundo tenham sido utilizados pelos melhoristas de plantas (CARTER et al., 2004). Um dos principais fatores para a subutilização dos recursos genéticos se dá devido à falta de caracterização destes acessos exóticos.

Apesar do sucesso de vários estudos envolvendo a caracterização fenotípica de acessos de banco de germoplasma (PERRY e MCINTOSH, 1991; SNELLER, 1994; GIZLICE; et al., 1994; BERNARD et al., 1998), os fenótipos podem ser altamente influenciados pela interação genótipo por ambiente (GxE), o que pode dificultar a avaliação de caracteres morfológicos, principalmente quando envolve a avaliação de caracteres quantitativos. Além disto, quase sempre os acessos são avaliados em condições de ambiente nos quais estes não possuem adaptação, dificultando ainda mais a sua caracterização.

Uma alternativa para minimizar tais problemas no estudo da diversidade genética é a utilização dos marcadores moleculares. A caracterização molecular foi primeiramente realizada em soja com a utilização de isoenzimas (Chen et al., 1989; Perry et al., 1991) e posteriormente com outros tipos de marcadores tais como RAPD (*Restriction fragment length polymorphism*), AFLP (*Amplified fragment length polymorphisms*) e SSR (*Simple sequence repeat*). Atualmente os SNPs (*Single nucleotide polymorphisms*) são as marcas mais utilizadas para diversos tipos de análises moleculares, incluindo estudos de diversidade genética, pois possuem uma série de vantagens, como alto nível de polimorfismo, acurácia, baixo custo por *data point*, além de permitirem a genotipagem em todo o genoma (VARSHNEY et al., 2009).

Hyten et al. (2006) sequenciaram com SNPs quatro populações de soja, representando grupos de genótipos antes e após eventos de gargalo genético (*bottlenecks*). Os autores observaram a perda de várias sequências raras e numerosas mudanças de frequências alélicas ao longo da história, além disso, chegaram à conclusão de que o gargalo genético com maior efeito na perda de variabilidade

genética foi a domesticação. Já Lam et al. (2010), utilizando análise de SNPs de alta performance, re-sequenciaram 17 linhagens de sojas selvagens e 14 cultivares comerciais, afim de verificar e comparar os padrões de variação. Como esperado os autores identificaram alta diversidade alélica nos genótipos selvagens e grande presença de desequilíbrio de ligação no genoma da soja como um todo.

O entendimento e a caracterização da diversidade genética de soja nos bancos de germoplasma ajudaria a melhor utilização dos recursos genéticos e a incorporação de novas fontes de variabilidade genética nos programas de melhoramento. Dado o exposto, o presente trabalho tem como objetivo a caracterização molecular da diversidade genética e estrutura populacional, em um painel com 80 acessos exóticos de soja e 15 cultivares comerciais utilizando marcadores SNPs distribuídos ao longo do genoma.

4.2 Material e Métodos

4.2.1 Material Vegetal

O painel de soja avaliado foi composto por 95 genótipos (ANEXO A), dentre estes 10 variedades comerciais brasileiras: IAC100, Conquista, CD 215, BMX Potência, A7002, Vmax, Sambaíba, Pintato e Paranagoiania, 2 linhagens do programa de melhoramento de soja da Universidade Estadual Paulista “Júlio de Mesquita” (UNESP-Jaboticabal): JAB 00-05-6/763D e JAB 00-02-2/2J3D, 4 linhagens do programa de melhoramento de soja da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ): LQ1050, LQ1505, LQ1421 e LQ1413. E 80 PI’s (*Plant Introductions*) de soja selecionadas em estudo prévio realizado por Mulato (2010) e que compreendem genótipos provindos de diferentes localidades do mundo.

4.2.2 Extração de DNA

Os 95 genótipos foram semeados em vasos na casa de vegetação. Tecido foliar foi coletado de uma planta a partir do segundo par de folhas verdadeiras, para extração do DNA genômico pelo método CTAB (DOYLE, DOYLE; 1990). O DNA total foi quantificado no aparelho QuantiFluor®, e as amostras foram diluídas para a

concentração de 100 ng μL^{-1} e acondicionadas em placas, para serem enviadas para empresa Affymetrix© na Califórnia Estados Unidos para genotipagem.

4.2.3 Genotipagem com marcadores SNPs

A genotipagem foi realizada na plataforma da Affymettrix (Axiom® Soybean Genotyping Array) contendo 186,961 mil marcadores SNPs para soja cultivada e selvagem. As sequências para o desenho do arranjo de SNPs foram fornecidas pela parceria entre a Affymettrix e os doutores Soon-Chun Jeong, Namshin Kim, e Jung-Kuyng Moon, do Instituto de Pesquisa de Biociências e Biotecnologia da Coréia (KRIBB).

4.2.4 Análise dos SNPs

Os SNPs foram pré-processados pelo software da Affymettrix, Axiom® Analysis Suite. Os seguintes filtros foram aplicados no número de SNPs originais: DQC $\geq$ 0,82 (*Dish quality control*, medida da resolução das distribuição dos valores de contraste), QC *call rate* (*Quality control call rate*) $\geq$ 92 , *Average call rate for passing* $\geq$ 97, Minor alele cutoff $\geq$ 2. Valores baseados na recomendação do software Axiom® Analysis Suite

Um número de 20 mil SNPs foi utilizado para que todas as 95 amostras passassem nos filtros mencionados anteriormente. Destes 20 mil SNPs, após a utilização dos filtros mencionados anteriormente 50,08% SNPs (10017) foram classificados como *PolyHighResolution*, classe esta recomendada para utilização pelo software e que apresenta, boa resolução dos *clusters* e ao menos dois exemplos de *minor allele*.

Estes 10017 SNPs foram utilizados para a análise de diversidade genética e estrutura de populações.

4.2.5 Análise de Diversidade

Para acessar a diversidade genética entre os acessos foi utilizado o software PowerMarker 3.25 (LIU e MUSE 2005), onde foram calculados a heterozigosidade

observada (Ho), a heterozigosidade esperada (He) ou diversidade do gene no caso de SNPs, porcentagem informativa de polimorfismo (PIC), a análise de variância molecular AMOVA, as estatísticas F_{st} para populações (WRIGHT, 1965), as frequências alélicas para o cálculo das distâncias Euclidianas e a construção do dendrograma pelo método *Neighbor Joining Tree* com a utilização de 1000 *bootstraps*. A obtenção do dendrograma consenso foi feita no programa PHYLIP 3.6 (FELSENSTEIN, 2005) e para edição e visualização do dendrograma final foi utilizado o software Mega 6 (TAMURA et al., 2013).

O cálculo da correlação cofenética para o dendrograma pelo método de *Neighbor Joining Tree,* foi feito utilizando o pacote do software R *adegenet* (JOMBART, 2008). Para fins de comparação também foi calculado este mesmo tipo de correlação utilizando agrumento do tipo UPGMA.

4.2.4 Análise de estrutura da população

Para as análises de estrutura da população foram utilizados dois softwares. Primeiro o pacote do R *adegenet* (JOMBART, 2008), empregando-se uma análise exploratória de dados, a Análise Discriminante de Componentes Principais (DAPC).

Na análise de DAPC os dados de marcadores foram primeiramente transformados pela análise de Componentes Principais (PCA), o número total de componentes foi mantido maximizando a quantidade de variância explicada. Para a definição do número *clusters* (*k*) foi feita pelo algoritmo *K means,* que identifica o melhor número de clusters a partir do melhor valor de BIC (*Bayesian Information Criterion*).

Utilizando os dados transformados para PCA e o número de clusters definido pelo algoritmo *k means*, foi deita a análise de DAPC. Foram retidos 40 componentes principais resultando na formação de uma função discriminante. Com as informações desta função discriminante foi gerado um gráfico do tipo *compoplot*, com as probabilidades de atribuição de cada individuo aos *clusters* formados.

O segundo software utilizado foi o STRUCTURE 2.3.4 (PRITCHARD et al., 2000) via abordagem Bayesiana, utilizando o modelo sem mistura (*no admixture*). As

frequências alélicas foram consideradas correlacionadas entre as populações, foram testados valores de *k* (número de populações prováveis) de 1 a 10, com dez repetições para cada *k*, um período de *burn-in* de 100 mil e dez mil repetições da cadeia de Markov (MCMC). A determinação do k mais provável foi feita de acordo com método proposto por Evanno et al. (2005) através da plataforma online STRUCTURE HARVESTER (EARL e VONHOLDT, 2012). A matriz Q do k mais provável foi então analisada pelo software CLUMPP (JAKOBSSON e ROSENBERG, 2007), os resultados foram inseridos no software online STRUCTURE PLOT (RAMASAMY et al., 2014) onde foi gerado o gráfico com as respectivas probabilidades de atribuição de cada individuo aos *clusters*.

4.3 Resultados e Discussão

Um total de 20 mil SNPs foram utilizados na genotipagem dos 95 acessos de soja, deste total apenas 10017 (50%) marcadores foram polimórficos. O conteúdo de informação de polimorfismo (PIC) para os marcadores, variou de 0,021 a 0,480, com uma média de 0,205 em todos os 95 genótipos analisados. Estes resultados demonstram a presença de diversidade genética entre os acessos. Os valores de PIC representam a diversidade gênica para um dado loco, tão logo, quanto mais elevados os valores de PIC maior a probabilidade de polimorfismo entre dois acessos para aquele loco (LI e NELSON, 2001). Hao et al. (2012) genotiparam 191 variedades crioulas de soja com 1536 SNPs e encontraram valores similares de PIC, variando de 0,20 a 0,45.

A média da heterozigosidade esperada (He), foi de 0,245, com valores variando de 0,028 a 0,5735. Enquanto que, a heterozigosidade observada (Ho) apresentou valores de 0 a 0,34, e uma média de 0,0723 nos 10017 locos. Os valores baixos de Ho são esperados visto que a soja é uma planta autógama e possui baixa quantidade de heterozigotos. Segundo Bai e Gai (2003), valores de heterozigosidade altos provavelmente são resultantes de polinização cruzada, a qual ocorre em torno de apenas 0,5% em soja, de forma natural.

Com o propósito de verificar distância genética entre os acessos de soja do painel, foi realizada a construção de um dendrograma com base nas distâncias Euclidianas pelo método *Neighbor Joining Tree* com a utilização de 1000 bootstraps. Este tipo de metogologia foi testada utilizando o cálculo de correlação cofenética, q qual é uma medida da acúracia do dendrograma em relação as distâncias originais. O valor obtido para o agrupamento do tipo foi *Neighbor Joining Tree* de 94,54% e para UPGMA o valor foi de 64,98% (Figura 1). Destacando a melhor precisão do tipo de agrupamento escolhido para os dados deste trabalho.

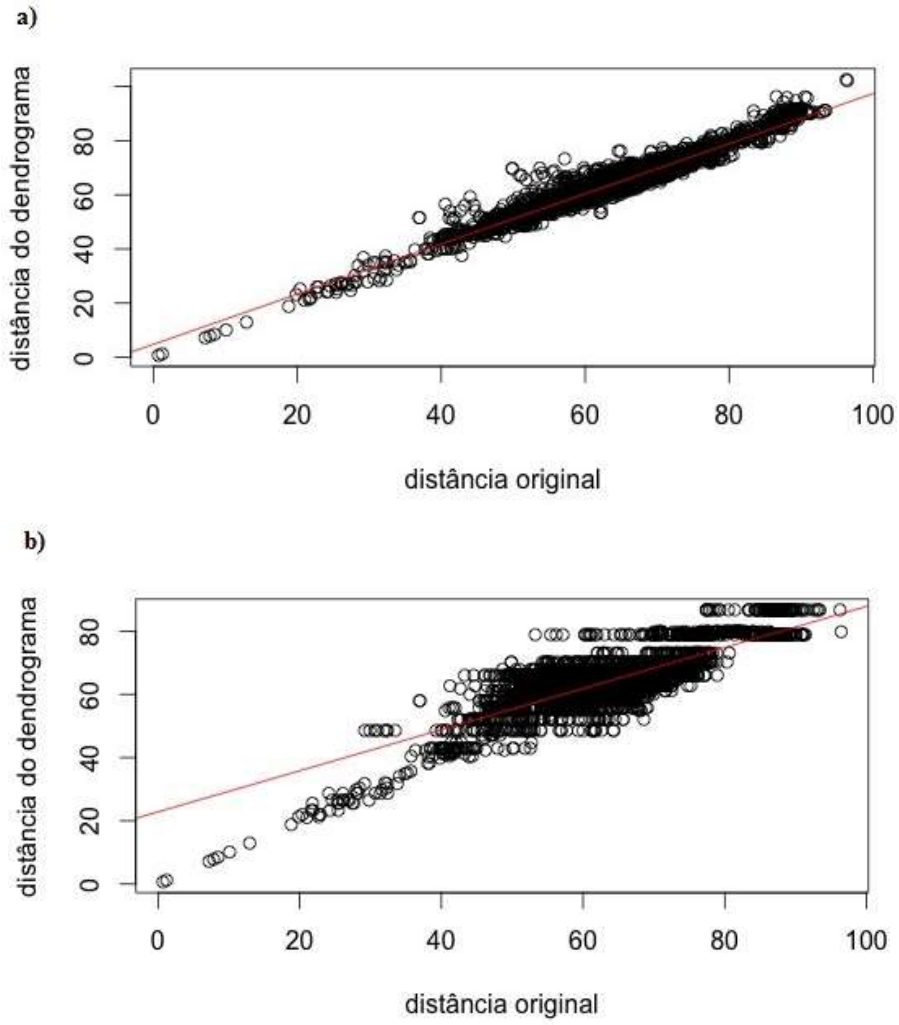

Figura 1 Correlação cofenética da matriz de distâncias Euclidianas baseada em Marcadores SNPs, a) correlação cofenética do método de *Neighbor Joining Tree,* b) correlação cofenética do método UPGMA

As distâncias genéticas variaram de 0,0001 entre os genótipos PI 265497 e PI 171451, a 0,6559 entre as linhagens Paranagoiania e PI 416828. Todos os 95 genótipos analisados foram classificados em dois grupos principais (Figura 2): o grupo I em azul, composto por 41 acessos, dentre eles está a testemunha brasileira Sambaíba, duas linhagens do programa de melhoramento genético da Escola Superior de Agricultura “Luiz de Queiroz” as LQ 1421 e LQ 1050, e a linhagem do programa de melhoramento genético da UNESP – JAB 00-02-2/2J3D Já o grupo II em em vermelho formado por 54 acessos, contendo a maioria das cultivares brasileiras utilizadas neste trabalho.

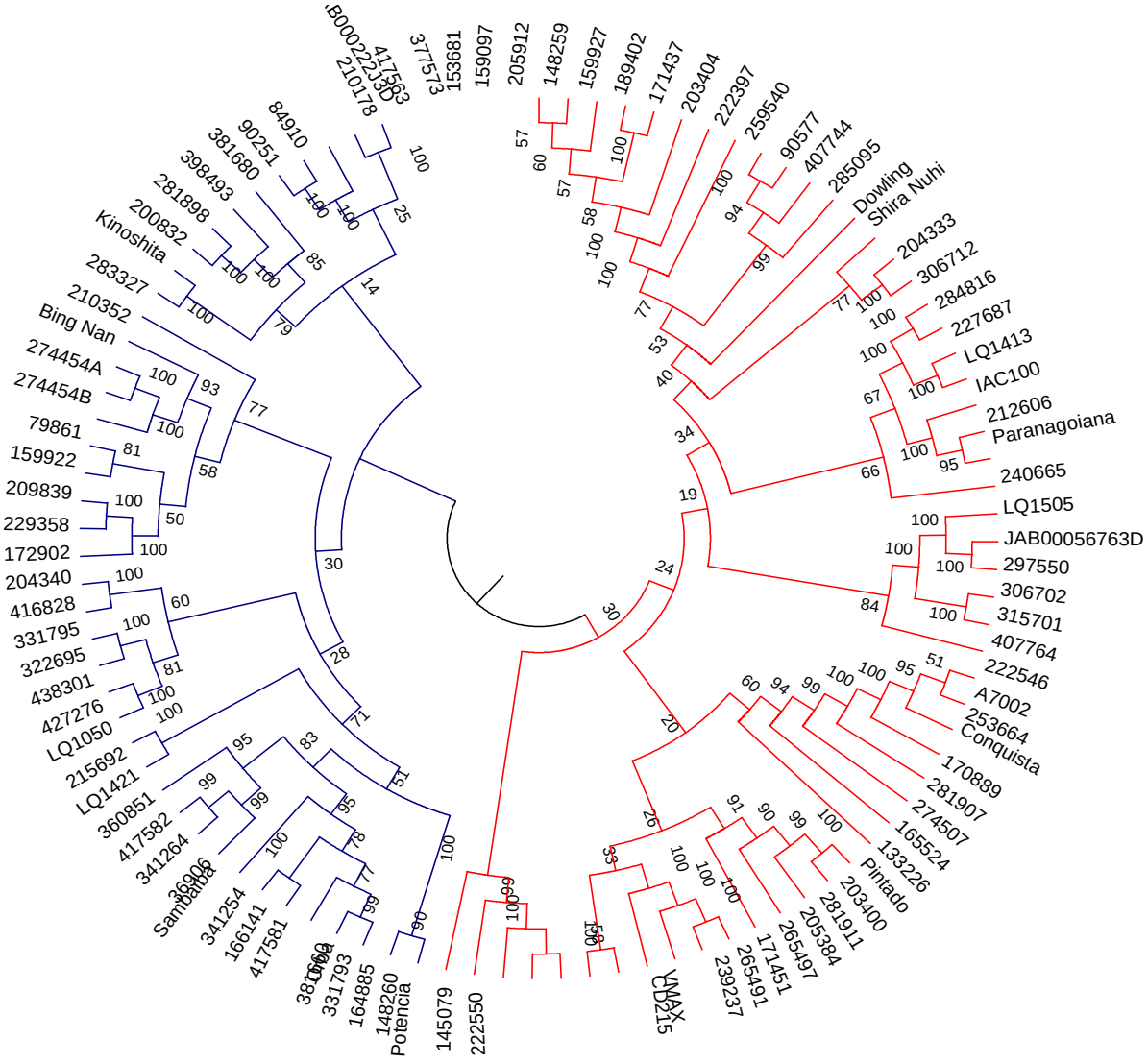

Figura 2 –Dendrograma via *Neighbor Joining Tree* com a utilização de 1000 bootstraps, utilizando distâncias Euclidianas obtido com dados de genotipagem de 95 linhagens de soja

Os acessos provenientes da China e Japão estiveram presentes em todos os dois grupos formados, uma explicação válida é que a região compreendida por estes países é considerada centro de diversidade e domesticação da soja e como tal, vários acessos provenientes destes países foram introgredidos ao redor do mundo como fonte de variabilidade e diversidade genética (LI e NELSON, 2001).

Em trabalho realizado por Li e Nelson (2001) os autores avaliaram a diversidade genética em três países considerados fontes de germoplasma para o mundo todo, China, Japão e Coréia do Sul. Utilizando marcadores RAPD em acessos de soja originários destes três países e técnicas de agrupamento hierárquicos, os autores foram capazes de identificar que a distância média entre acessos da China é muito maior que dos acessos do Japão e Coréia do Sul, entretanto é menor quando comparada com as distâncias entre os três países. Os acessos chineses foram completamente separados dos demais acessos, contudo a análise de agrupamento não foi capaz de separar entre os genótipos de origem japonesa e sul coreana, indicando similaridade genética entre estes. Tal relação também pode ser observada pela formação do grupo II (Figura 2).

Dentro destes dois grupos do dendrograma (Figura 2), os acessos também foram divididos em subgrupos que possuem padrões semelhantes. As cultivares comerciais Vmax, CD 215 foram alocadas juntas, indicando grau de proximidade genética entre estes três genótipos, o mesmo pode ser observado para as linhagens Conquista e A7002, assim como para o subgrupo Paranagoiania e IAC 100. Pode-se observar também no grupo II a alocação de duas linhagens descritas na literatura como resistentes a afídeos e percevejos (MICHEREFF et al., 2014), respectivamente, as linhagens Dowling e IAC 100, indicativo de uma provável similaridade genética entre estes dois genótipos.

Os dois grupos formados no dendrograma (Figura 2) apresentam padrões de germoplasma distintos. O grupo em vermelho contém grande parte das cultivares comerciais brasileiras, tais como Vmax e Potencia, assim como linhagens dos EUA, estes indivíduos representam basicamente o germoplasma americano que foi introgredido no sul do Brasil para inicio do cultivo da soja no país. Na literatura alguns

trabalhos já demonstraram que o germoplasma e a base genética da soja no Brasil foi formada principalmente por linhagens e recursos genéticos introduzidos da América do Norte. Segundo trabalho realizado por Paludzyszyn Filho et al. (1993) basicamente as introgressões de soja brasileiras foram de cultivares originárias do sul dos Estados Unidos, mesma conclusão foi feita por Hiromoto e Vello, (1986).

O grupo representado pela cor azul (Figura 2) representa muito bem o germoplasma asiático. Comparando a base genética do Brasil com a Japonesa e Chinesa, Wysmierski e Vello (2013) relataram que apenas nove e sete ancestrais, respectivamente, são compartilhados entre os referidos países (Japão e China) e o germoplasma brasileiro.

Mulato et al (2010) analisando os mesmos acessos exóticos deste trabalho utilizando marcadores microsatélites e presentes em regiões genômicas e expressas encontraram resultados diferentes no dendrograma pelo método UPGMA baseado na distância de Rogers-W. O dendrograma obtido apresentou 5 grupos com um ponto de corte de 0,82.

Afim de verificar melhor a estrutura populacional do painel de acessos de soja foram feitas também as análises de estruturação genética da população. Utilizando o software STRUCTURE (abordagem Bayesiana) e uma metodologia mais simples e rápida utilizando análise exploratória de dados, a análise Discriminante via Componentes Principais (DAPC), que vem sendo adotada principalmente para dados com milhares de marcadores (JOMBART et al., 2010).

A análise de DAPC sem informação a *priori* de populações ou grupos obteve um número ótimo de grupos (*k*) igual a 2 correspondendo ao menor valor de BIC (659,01), o qual refere-se ao critério de informação Bayesiano, quanto menor o valor de BIC maior a acurácia do modelo escolhido (JOMBART et al., 2010). Foram retidos 40 componentes principais e uma função discriminante, correspondendo a 82% da variação total presentes nos dados analisados. Grande parte dos padrões comerciais foram alocados no grupo 2 (Figura 3), com exceção da cultivar Sambaíba e das linhagens da ESALQ LQ 1050 e LQ 1421. Este mesmo padrão de agrupamento dos genótipos brasileiros foi detectado na análise *Neighbor Joining Tree* (Figura 2).

Na Figura 2 observa-se o gráfico da probabilidade de atribuição de cada um dos 95 acessos aos 2 grupos formados pela análise de DAPC. Entre os acessos analisados apenas a cultivar americana Downling apresentou probabilidade de atribuição aos grupos inferior a 100%.

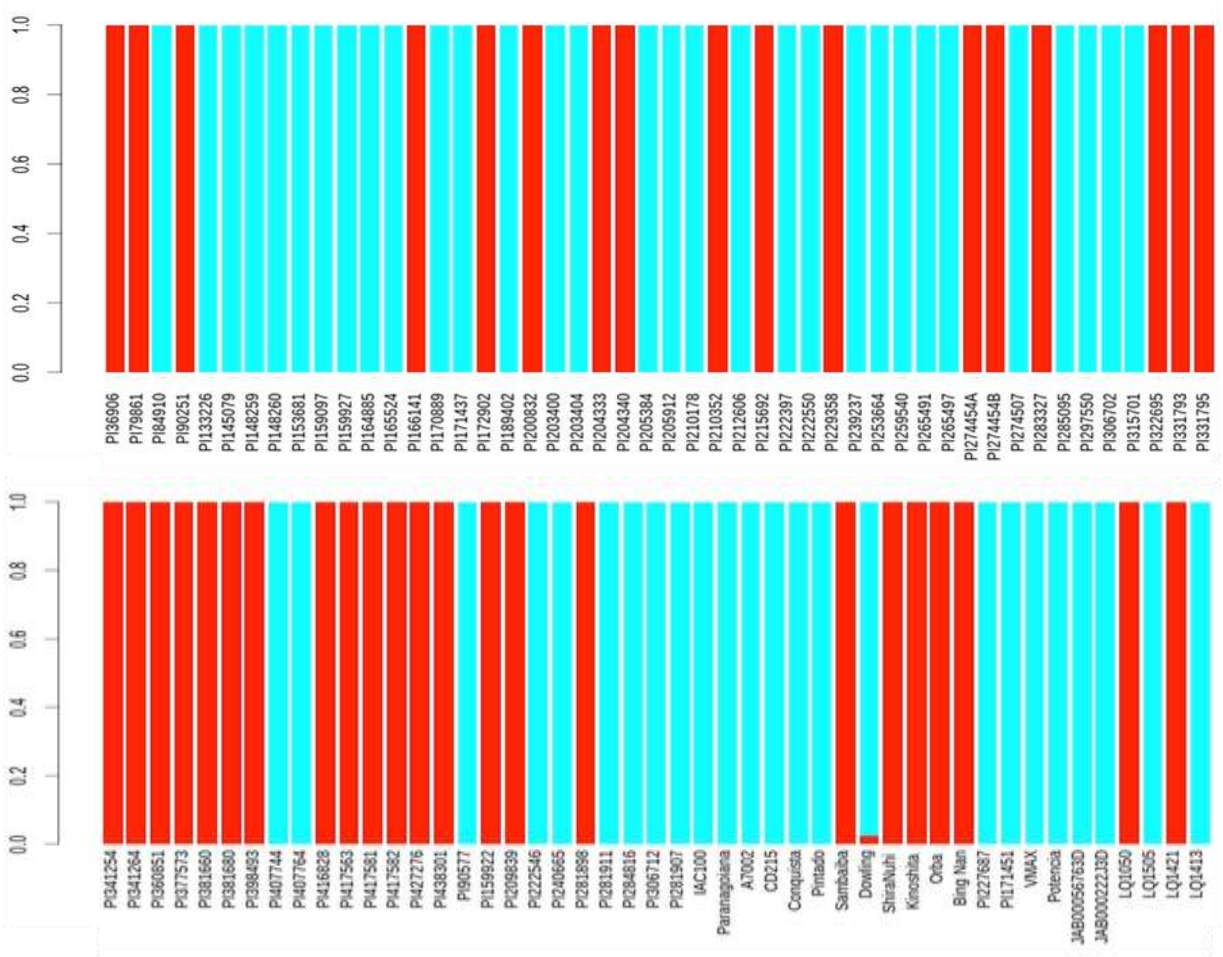

Figura 3 – Gráfico da probabilidade de atribuição dos 95 linhagens de soja aos 2 grupos (k=2) de acordo com a análise discriminante por componentes principais (DAPC). Cada grupo representado por uma cor: k=I em vermelho e k=II em azul.

O valor médio de F_{st} obtido para estas duas populações (Figura 3) foi de 0,0949. O F_{st} estimado pelo método proposto por Wright (1965) pode variar de 0 a 1, onde 0 significa que as populações tem frequência de alelos idênticas, e 1 as populações fixaram alelos diferentes. Sendo assim, os valores obtidos acima indicam certo grau de compartilhamento genético e endogamia entre os dois *clusters*.

Pela análise de variância molecular (AMOVA) (Tabela 1) a maior porcentagem da variação total presente na análise de DAPC foi atribuída dentro de populações, aqui consideradas como grupos ou *clusters* (Figura 2) . O grupo I (vermelho) contando com 37,88% da variação presente enquanto que o grupo II apresentou 41,26% da variabilidade total.

Tabela 1. Análise de Variância Molecular (AMOVA) para 10017 marcadores SNPs utilizando como populações os *clusters (k)* formados pela análise de DAPC

Fonte de Variação	**Soma de Quadrados**	**Porcentagem Explicada**
Entre clusters	26010,37232	57,706272
Dentro de k I	170761,6	37,8849452
Dentro de k II	185969,3833	41,2589241
dentro de individuos k I	13613	3,0201624
dentro de individuos k II	54383	12,0653412
Total	450737,3556	1

O mesmo tipo de análise de estrutura populacional foi feita, utilizando uma abordagem Bayesiana, com o software STRUCTURE. O número de grupos que melhor descreveu o painel de soja analisado foi k=2, resultado semelhante ao encontrado pela análise de DAPC. Entretanto, a atribuição de alguns genótipos aos grupos foi diferente nos dois tipos de abordagem , em ambas as análises a presença de diferentes cores no gráfico em um mesmo indivíduo indica a porcentagem do genoma compartilhado dentro de cada grupo.

Em trabalho realizado por Sigrist (2012) utilizando um grupo similar de acessos de soja analisado neste estudo e marcadores, o autor encontrou na análise de estrutura da população um número de clusters *k=2*. As cultivares brasileiras foram agrupadas em um mesmo *cluster* e os acessos exóticos apresentaram um padrão de agrupamento que não reflete o país de origem dos mesmos. Tal resultado similar, corrobora as análises realizadas e traz confiabilidade ao trabalho.

O valor médio para estatística F_{st} para os dois clusters encontrados pela análise Bayesiana, foi 0,08424 valor próximo ao encontrado entre os grupos definidos pela análise de DAPC. Para os resultados da AMOVA (Tabela 2) a maior parte da variação totalexplicada foi presente dentro dos clusters e não entre estes, para o grupo k I o

percentual de variação explicado foi de 50,48% enquanto que, no grupo II este valor foi de 29,26%. Sendo assim, o grupo I concentrou a maior parte da variabilidade presente nas duas populações identificadas pelo software STRUCTURE.

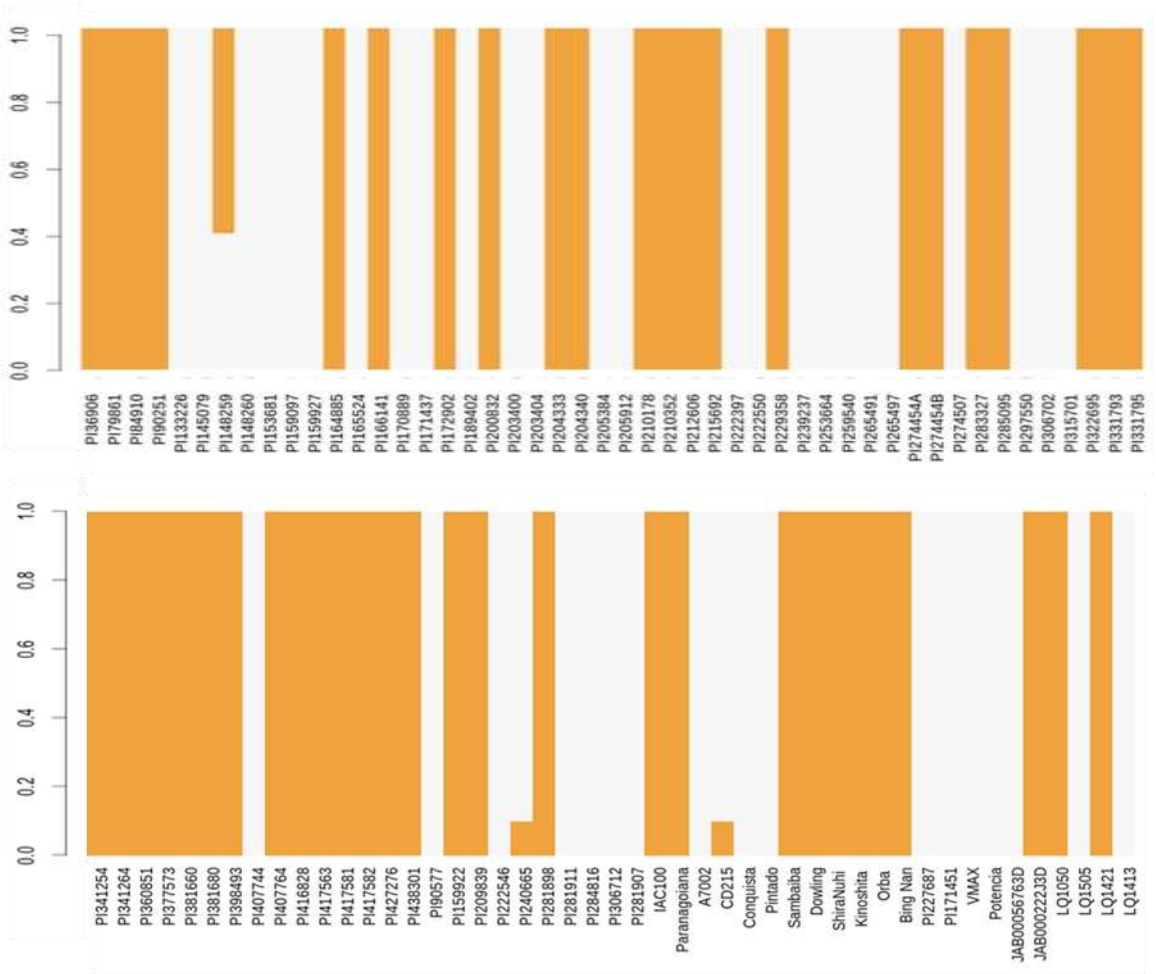

Figura 4 – Gráfico da probabilidade de atribuição dos 95 linhagens de soja aos 2 grupos (k=2) de acordo com a análise Bayesiana pelo software STRUCTURE. Cada grupo representado por uma cor: k=I em laranja e k=II em branco.

Observa-se que as linhagens alocadas aos grupos pela análise de DAPC (Figura 3) foram mais semelhantes ao agrupamento por vizinho mais próximo (Figura 2). Pela análise de DAPC, as cultivares brasileiras foram alocadas em um só grupo, com exceção das linhagens Sambaíba, LQ 1050 e LQ 1421. Já na análise pelo STRUCTURE (Figura 4) estas testemunhas tiveram distribuição igualitária em dois grupos. No primeiro grupo, em laranja, ficaram as cultivares IAC 100, Paranagoiania, Sambaíba, JAB000222J3D, LQ 1050 e LQ 1421, e o segundo grupo identificado pela

cor branca no gráfico, as linhagens A7002, CD215, Conquista, Pintado, Vmáx, Potência, JAB00056763D, LQ 1505 e LQ 1413.

Tabela 2. Análise de Variância Molecular (AMOVA) para 10017 marcadores SNPs utilizando como populações os *clusters (k)* formados pelo STRUCTURE

Fonte de Variação	**Soma de Quadrados**	**Porcentagem Explicada**
Entre clusters	23318,22028	51,733498
Dentro de k I	227551,2277	50,4842177
Dentro de k II	131871,9076	29,2569289
dentro de individuos k I	26427	58,630596
dentro de individuos k II	41569	9,222444
Total	450737,3556	1

A estrutura populacional é uma consequência dos efeitos de fatores evolutivos tais como deriva genética, migração, mutação e seleção, mas principalmente do tipo de sistema reprodutivo. Espécies autógamas, como a soja, são propensas a apresentar menor diversidade alélica, altos coeficientes de endogamia e consequentemente baixa heterozigosidade, além de possuírem alta diferenciação entre diferentes tipos de populações quando comparadas a espécies alógamas (WRIGHT et al., 1921).

Neste estudo não houve a associação entre a região geográfica de coleta destes materiais e os clusters obtidos pelas três técnicas aqui apresentadas, os genótipos analisados provém de 37 países ao redor do mundo, entretanto foram detectados de 2 a 3 clusters nas análises de agrupamento e estrutura da população. Uma das prováveis causas da baixa estratificação em grupos seja a utilização de genótipos já melhorados pelo homem, pois apesar de serem genótipos exóticos às condições brasileiras, muitos dos acessos apresentados são cultivares comerciais em seus países de origem. Na cultura da soja trabalhos envolvendo a utilização destes dois tipos de metodologia são inexistentes, principalmente devido a DAPC ser uma análise relativamente recente (JOMBART et al., 2010). Entretanto, vários trabalhos podem ser encontrados na literatura comparando este dois tipos de análise na identificação da estrutura de populações (MORGAN et al., 2013; POMETTI, et al., 2014), é possível constatar nestes estudos a similaridade entre estes dois métodos e a acurácia ao identificar o número de populações. Segundo trabalho de Jombart et al.

(2010), no qual este implementa a técnica de DAPC, esta metodologia permite maior rapidez na análise de grandes conjuntos de dados, fator de contraponto a técnica de agrupamento empregada pelo STRUCTURE, visto que, os algoritmos deste software requerem grande demanda computacional. No mesmo estudo as duas técnicas se mostraram acuradamente similares na detecção do número de clusters em dados simulados.

4.4 Conclusões

A caracterização molecular indicou a presença de diversidade genética no painel de acessos avaliados, semelhante a encontrada na literatura em estudos similares. Foi possível detectar a presença de dois grupos nas duas análises de estruturação genética da população utilizadas, assim como no método de agrupamento via Neighbor Joining Tree. além de confirmar resultados obtidos anteriormente isto permitiu maior confiabilidade nas análises realizadas.

A análise de DAPC se mostrou eficaz e pode ser utilizada quando há uma grande quantidade de dados como aqueles de genotipagem presentes ao longo de todo o genoma.

Referências

BAI, Y. N.; GAI, J. Y. Development of soybean cytoplasmic-nuclear male-sterile line NJCMS2A and restorability of its male fertility. **Scientia Agricultura Sinica**, Beijing, v. 36, n. 7, p. 740-745, 2003.

BERNARD, R. L.; CREMEENS, C. R.; COOPER, F. I.; COLLINS, O. A. **Evaluation of the USDA Soybean Germplasm Collection: Maturity Groups 000 to IV (FC 01.547 to PI 266.807)**. Washington: U.s. Department Of Agriculture, 1998. (Technical Bulletin n. 1844).

CARTER, T. E.; NELSON, R. L.; SNELLER, C. H.; CUI, Z. Genetic Diversity in Soybean. In: BOERMA, H. R. (Ed.); SPECHT, J. E. Soybean: **improvement, production and uses**. 3. ed. Madison: American Society Of Agronomy, 2004. Cap. 8. p. 303-396.

CHEN, Z.-L.; NAITO, S.; NAKAMURA, I.; BEACHY, R. N. Regulated expression of genes encoding soybean β-conglycinins in transgenic plants. **Developmental genetics**, New York, v. 10, n. 2, p.112-122, 1989.

EARL, D. A.; VONHOLDT, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, London, v. 4, n. 2, p.359-361, 2011.

EVANNO, G.; REGNAUT, S.; GOUDET, J.. Detecting the number of clusters of individuals using the software structure: a simulation study. **Molecular Ecology**, Oxford, v. 14, n. 8, p.2611-2620, jul. 2005.

FELSENSTEIN, J. **PHYLIP (Phylogeny Inference Package)** version 3.6. Department Of Genome Sciences, University Of Washington, Seattle, 2005.

GIZLICE, Z.; CARTER, T. E.; BURTON, J. W. Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. **Crop Science Society Of America**, Madison, v. 34, n. 5, p.1143-1151, 1994.

HIROMOTO, D. M.; VELLO, N. A.. The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. **Brazilian Journal Of Genetics,** Ribeirão Preto, v. 09, n. 2, p.295-306, 1986.

HYTEN, D. L.; SONG, Q.; ZHU, Y. Impacts of genetic bottlenecks on soybean genome diversity. **Proceedings Of The National Academy Of Sciences**, Washington, v. 103, n. 45, p.16666-16671, 2006.

JOMBART, T.. Adegenet: a R package for the multivariate analysis of genetic markers. **Bioinformatics**, Oxford, v. 24, n. 11, p.1403-1405, 2008.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **Bmc Genetics**, Londres, v. 11, n. 1, p.94-94, 2010.

JAKOBSSON, M.; ROSENBERG, N. A.. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics**, Oxford, v. 23, n. 14, p.1801-1806, 2007.

LAM, H.; XU, X.; LIU, X.; CHEN, W.; YANG, G.; WONG, F.; LI, M.; HE, W.; QIN, N.; WANG, B.; LI, J.; JIAN, M.; 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**, Nova York, v. 42, n. 12, p.1053-1059, 2010.

LI, Z.; NELSON, R. L. Genetic Diversity among Soybean Accessions from Three Countries Measured by RAPDs. **Crop Science Society Of America**, Madison, v. 41, n. 4, p.1337-1347, 2001.

LIU, K.; MUSE, S.V. PowerMarker: an integrated analysis environment for genetic marker analysis. **Bioinformatics**, Oxford, v.21, p.2128-2129, 2005.

MICHEREFF, M. F. F.; MICHEREFF FILHO, M.; BLASSIOLI-MORAES, M. C.; LAUMANN, R. A.; DINIZ, I. R.; BORGES, M. Effect of resistant and susceptible soybean cultivars on the attraction of egg parasitoids under field conditions. **Journal Of Applied Entomology**, Berlin, v. 139, n. 3, p.207-216, 21 jul. 2014.

MORGAN, E. M. J.; GREEN, B. S.; MURPHY, N. P.; STRUGNELL, J. M. Investigation of Genetic Structure between Deep and Shallow Populations of the Southern Rock Lobster, Jasus edwardsii in Tasmania, Australia. Plos One, San Francisco, v. 8, n. 10, p. e77978, 18 out. 2013.

PALUDZYSZYN FILHO, E.; KIIHL, R. A. S.; ALMEIDA, L.A. Desenvolvimento de cultivares de soja na região Norte e Nordeste do Brasil. In: SIMPÓSIO SOBRE CULTURA DA SOJA NOS CERRADOS, 0., 1992, Uberaba. **Anais**. Piracicaba: Potafos, 1993. v. 0, p. 255 - 265.

PERRY, M. C.; MCINTOSH, M. S. Geographical patterns of variation in the USDA soybean germplasm collection: I. Morphological traits. **Crop Science Society Of America**, Madison, v. 31, n. 5, p. 1350-1355, 1991.

PERRY, S. E.; SUMMAR, M. L.; PHILLIPS, J. A. Linkage analysis of the human dopamine β-hydroxylase gene. **Genomics**, San Diego, v. 10, n. 2, p.493-495, 1991.

POMETTI, C. L.; BESSEGA, C. F.; SAIDMAN, B. O.; VILARDI, J. C. Analysis of genetic population structure in *Acacia caven* (Leguminosae, Mimosoideae), comparing one exploratory and two Bayesian-model-based methods. **Genetics And Molecular Biology**, Ribeirão Preto, v. 37, n. 1, p.64-72, 2014.

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

RAMASAMY, R.; RAMASAMY, S.; BINDROO, B.; NAIK, V. G. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springer **Plus**, Heidelberg, v. 3, n. 1, p.431-433, 2014.

SNELLER, C. H. Pedigree Analysis of Elite Soybean Lines. **Crop Science Society Of America**, Madison, v. 34, n. 6, p.1515-1522, 1994.

TAMURA, K; PETERSON, D; PETERSON, N; STECHER, G; NEI, M; KUMAR, S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. **Molecular Biology And Evolution**, Chicago, v. 28, n. 10, p.2731-2739, 2011.

VARSHNEY, R.K.; NAYAK, S.N.; MAY, G.D.;JACKSON, S.A. Next generation sequencing technologies and their implications for crop genetics and breeding. **Trends Biotechnology**, Amsterdam, v.27, p. 522–530, 2009.

WRIGHT, S. Systems of mating. **Genetics**, Austin, v. 6, p.111-178, 1921.

WRIGHT, S. The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution,** Malden, v. 19, p.395-420, 1965.

WYSMIERSKI, P. T.; VELLO, N. A. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. **Genetics And Molecular Biology**, Ribeirão Preto, v. 36, n. 4, p.547-555, 2013.

5 MAPEAMENTO ASSOCIATIVO PARA PRODUTIVIDADE DE GRÃOS EM PAINEL DE ACESSOS DE SOJA

Resumo

O mapeamento associativo vem nos últimos anos ganhando destaque e sendo cada vez mais utilizado no melhoramento vegetal, visando à identificação de regiões do genoma associadas a características de interesse. O objetivo deste trabalho é identificar regiões do genoma associadas a produtividade de grãos em um painel de acessos de soja via mapeamento associativo. O painel de acessos foi composto por 95 indivíduos, dentre estes 80 acessos exóticos e 15 cultivares brasileiras. Os acessos foram fenotipados nos anos agrícolas de 2012/2013 e 2013/2014, nas cidades de Piracicaba-SP, Jaboticabal-SP e Ponta Grossa-PR, totalizando cinco ambientes. O delineamento utilizado foi um Alpha Látice 5x19, com parcela de quatro linhas de cinco metros e três repetições. A genotipagem foi realizada através do Axiom® Soybean Genotyping Array contendo 10017 SNPs polimórficos para os acessos genotipados. A análise dos dados fenotípicos foi feita no software SELEGEN utilizando modelos mistos. Já a análise de associação foi efetuada pelo software TASSEL, utilizando o modelo misto MLM. Duas abordagens foram utilizadas na análise de associação, a primeira utilizando as médias fenotípicas ajustadas para BLUP dos cinco ambientes e a segunda utilizando apenas as médias ajustadas para cada local individualmente. Foram detectadas sete associações marcador-característica com $p<0,001$ e com correção para múltiplos testes $q<0,1$. Dentre estas, quatro estão presentes tanto no modelo da análise conjunta dos cinco ambientes quanto para o ambiente dois. Os demais marcadores foram significativos somente para este último local, o qual foi o único ambiente a apresentar associações significativas.

Palavras chave: *Glycine max;* Germoplasma; Modelos Mistos; TASSEL

Abstract

The associative mapping has gained prominence in recent years, and have been increasingly used in plant breeding, aiming the identification of genomic regions associated with features of interest. The objective of this study is to identify regions of the genome associated with grain yield in a panel of soybean lines through association mapping. The panel was composed of 95 individuals, among them 80 exotic lines and 15 Brazilian cultivars. The accessions were phenotypes in the agricultural season of 2012/2013 and 2013/2014, in Piracicaba-SP, Jaboticabal-SP and Ponta Grossa-PR, in five environments. The design was an alpha lattice 5x19, with plots of four rows of five meters and three replications. Genotyping was performed by Axiom® Soybean Genotyping Array containing 10017 polymorphic SNPs for the genotyped lines. The analysis of phenotypic data was made in SELEGEN software using mixed models. The association analysis was performed by TASSEL software using the mixed model MLM. Two approaches were used in the association analysis, the first using phenotypic average adjusted to BLUP values for the five different environments and the second

one using only the means for each enviroment individually. Seven marker-trait associations were detected with p <0.001 and with correction for multiple tests q <0.1. Among these, four are present both in in the model of joint analysis of the five environments as well at the environment two. The other markers were significant only for the latter site, which was the only enviroment to show significant associations.

Key words: *Glycine max*; Germplasm; Mixed models; TASSEL

5.1 Introdução

A produtividade de grãos é o principal objetivo de ampla maioria dos programas de melhoramento genético de plantas. Não obstante, na cultura da soja o mesmo também é verdade. Incrementos na produtividade de grãos desta oleaginosa se fazem cada vez mais necessários, principalmente devido a demanda crescente populacional e as limitações de áreas agricultáveis (BEDDINGTON, 2010).

Entretanto a produtividade não é um caráter facilmente mensurável, devido a sua natureza quantitativa , no qual há um grande número de genes envolvidos no controle e grande influência ambiental (ALLARD, 1999; RAMALHO et al., 2008). Devido a isto a avaliação e estudo desta característica, utilizando somente dados fenotípicos pode demandar tempo e recursos que muitas vezes o melhorista não possui.

Umas das ferramentas que visa auxiliar o melhoramento de plantas são os marcadores moleculares. Estes podem ser empregados em três abordagens no melhoramento de plantas: detecção e mapeamento de QTLs (*Quantitative trait loci*), a seleção assistida por marcadores (MAS) e por último a seleção genômica ampla (GWS) (RESENDE et al., 2013).

As técnicas de mapeamento visam identificar associações entre os alelos dos marcadores e as variações fenotípicas dos caracteres quantitativos. Existem basicamente duas abordagens para a detecção de QTLs, o mapeamento via análise de ligação e o mapeamento pela análise do desequilíbrio de ligação, ou mapeamento associativo (GWAS) (RESENDE et al., 2013).

O mapeamento associativo busca correlações significativas entre um loco marcador e o fenótipo da característica de interesse (GUPTA et al., 2005). Ambos os tipos de mapeamento estão baseados no desequilíbrio de ligação entre marcador e um

dado loco para a característica de interesse, entretanto o mapeamento associativo utiliza populações naturais, contando assim com várias gerações de recombinação e detectando apenas as associações marcador-característica fortemente ligadas.

Além disso, a análise de associação oferece vantagens em relação ao mapeamento de ligação, como maior resolução do mapa, maior número de alelos e menor gasto de tempo visto que não é necessário cruzamentos específicos para a geração da população a ser mapeada (FLINT-GARCIA et al., 2003).

Na cultura da soja vários estudos envolvendo GWAS foram realizados nos últimos anos envolvendo características como teor de óleo e proteína, grupo de maturação, altura de planta e florescimento, clorose devido à deficiência de ferro. (HWANG et al., 2014; MAMIDI et al., 2014; ZHANG et al., 2015)

Sigrist (2012) utilizando 114 marcadores microsatélites em 89 linhagens de soja entre elas cultivares brasileiras e *Plant Introductions* de diversos países, realizou o mapeamento associativo para produtividade de grãos e características correlacionadas em soja. O autor encontrou 285 associações significativas, dentre estas 30% das associações já descritas previamente na literatura.

Utilizando linhagens similares ao estudo de Sigrist (2010), este estudo tem como objetivo o mapeamento associativo para detecção de regiões do genoma associadas a produtividade de grãos em um painel de acessos de soja provenientes de diferentes partes do mundo, genotipados com marcadores SNPs e avaliados fenotipicamente em cinco ambientes.

5.2 Material e Métodos

5.2.1 Material Vegetal

O painel de acessos de soja foi composto por 80 genótipos exóticos de soja de diversos países do mundo e 15 cultivares brasileiras. As origens de cada acesso e testemunha podem ser observadas no ANEXO A:

5.2.2 Fenotipagem

Os experimentos de campo foram conduzidos nos anos agrícolas de 2012/2013 nas cidades de Piracicaba – São Paulo (SP) (Ambiente 1), Jaboticabal – SP (Ambiente 2) e Ponta Grossa – Paraná (Ambiente 3), e em 2013/2014 nas cidades de Piracicaba (Ambiente 4) e Jaboticabal (Ambiente 5). Os experimentos foram conduzidos no delineamento Alfa-Látice 5x19, com três repetições e parcela experimental de 4 linhas de 5 metros, com espaçamento entre linhas de 0,5 m, sendo colhidas apenas as 2 linhas centrais da parcela.

A característica avaliada foi a produtividade de grãos em kg ha^{-1}, mensurada através de pesagem após a colheita, secagem e limpeza dos grãos.

5.2.3 Genotipagem com marcadores SNPs

Folhas das 95 linhagens foram coletadas em casa de vegetação, após o aparecimento do primeiro par de folhas verdadeiras. As folhas foram maceradas com macerador automático e após procedeu-se a extração de DNA pelo protocolo CTAB (DOYLE, 1990). A qualidade do DNA foi avaliada por eletroforese em gel de agarose a 1% corado com SYBRSafe (Invitrogen). A quantificação foi feita no aparelho Quantifluor®, e as amostras foram diluídas para a concentração de 100 ng μL^{-1} e acondicionadas em placas e gelo seco, para serem enviadas para empre Affymetrix®, em Santa Clara Califórnia, nos Estados Unidos, para genotipagem com marcadores SNPs (*Single Nucleotide Polymorphism*).

A plataforma utilizada para a genotipagem foi a Axiom® Soybean Genotyping Array contendo 186,961 SNPs mapeados com base no genoma de referência Williams 82.

5.2.4 Análise dos SNPs

Os SNPs foram pré-processados pelo software da Affymettrix, Axiom® Analysis Suite. Os seguintes filtros foram aplicados no número de SNPs originais: DQC $\geq$ 0,82 (*Dish quality control*, medida da resolução das distribuição dos valores de contraste),

QC *call rate* (*Quality control call rate*) $\geq$ 92 , *Average call rate for passing* $\geq$ 97, Minor alele cutoff $\geq$ 2. Valores baseados na recomendação do software Axiom® Analysis Suite

Um número de 20 mil SNPs foi utilizado para que todas as 95 amostras pasassem nos filtros mencionados anteriormente. Destes 20 mil SNPs, após a utilização dos filtros mencionados anteriormente 50,08% SNPs (10017) foram classificados como *PolyHighResolution*, classe esta recomendada para utilização pelo software e que apresenta, boa resolução dos *clusters* e ao menos dois exemplos de *minor allele*.

Os 10017 SNPs filtrados foram utilizados para as análises de estrutura de população, desequilíbrio de ligação, matriz de parentesco e mapeamento associativo.

5.2.4 Análise de dados fenotípicos

Os dados fenotípicos foram analisados pelo software SELEGEN (RESENDE et al., 1994) utilizando modelos mistos, considerando o efeito de genótipos, blocos e interação como de efeito aleatório. Primeiro foram feitas as análises individuais para cada um dos cinco ambientes e posteriormente as análises conjuntas.

As médias corrigidas para valores de BLUP (*Best Linear Unbiased Predictions*) para os cincos ambientes em conjunto assim como as médias ajustadas para cada ambiente individualmente foram utilizadas no mapeamento.

5.2.5 Estrutura de populações

A estrutura de população (matriz Q) foi inferida via abordagem Bayesiana no software STRUCTURE 2.3.4 (PRITCHARD et al., 2000) a partir dos 10017 SNPs obtidos anteriormente. O modelo utilizado foi o de não mistura (*no-admixture*) e frequências alélicas correlacionadas entre as populações. Também foi utilizado um número de subpopulações hipotéticos (*k*) de 1 a 10, com dez repetições para cada *k*, um período de *burn-in* de 100 mil e dez mil repetições da cadeia de Markov (MCMC) e um O valor de *k* mais provável foi determinado pela método de Evanno et al. (2005) na

plataforma online STRUCTURE HARVEST. A matriz Q do k mais provável foi então analisada pelo software CLUMPP (JAKOBSSON e ROSENBERG, 2007). O melhor k para este painel de acessos foi *k=2*. Os dados referentes ao número de subpopulações *k=2* foram formatados para utilização como matriz Q no software TASSEL 5.0 (BRADBURY et al., 2007).

5.2.6 *Matriz de parentesco*

A matriz de parentesco (k) foi inferida pelo software TASSEL 5.0, utilizando os mesmos 10017 SNPs. O calculo foi feito pela opção "*scaled IBS*" método desenvolvido por Endelman e Jannink (2012). Nesta metodologia os genótipos são codificados como 2, 1 ou 0, e cada número corresponde a contagem de um dos alelos para o loco em questão. Os dados perdidos foram então substituídos pela média do *score* genotípico para aqueles locos.

5.2.7 Desequilibrio de Ligação e MAF

Os 10017 SNPs foram filtrados para um MAF (*Minimum Minor Allele Frequency*) maior que 0,005, de acordo com a recomendação do software. Destes restaram 4992 SNPs que foram utilizados no cálculo do desequilibrio de ligação (DL). O DL entre pares de marcadores foi calculado pelo coeficiente de determinação *r* elevado ao quadrado, utilizando o teste de permutação rápida do software TASSEL 5.0. Os cálculos foram feitos para cada grupo de ligação separadamente, evitando-se assim o desequilibrio devido a outros fatores que não a ligação entre marcadores. Os DL foram considerados significativos quando $p<0,01$.

Os valores de r^2 significativos e as respectivas posições em pares de base (pb) foram então plotados em um gráfico utilizando o programa Excel.

5.2.7 Análise de Associação

A análise de associação foi feita utilizando o software TASSEL 5.0 com o modelo linear misto (MLM), que usa como covariáveis para correção dos efeitos de sub-estruturação da população as matrizes *Q* (estrutura da população) e *k* (matriz de parentesco). Dois tipos de médias fenotípicas foram utilizadas, primeiro as médias conjuntas dos cinco ambientes analisados, corrigidas para valores de BLUP e, segundo, o modelo considerando as médias ajustadas para cada ambiente individualmente. O modelo utilizado MLM pode ser descrito abaixo:

$$y = X\beta + Sa + Qv + Zu + e$$

$y =$ vetor de observações fenotípicas;

$\beta =$ vetor de efeitos fixos (demais efeitos excluindo-se os de marcadores e estrutura da população);

$a =$ vetor de efeitos fixos de marcadores;

$v =$ vetor de efeitos fixos da estrutura da população;

$u =$ vetor de efeitos aleatórios poligênicos desconhecidos;

$e =$ vetor de efeitos aleatórios residuais;

$Q =$ matriz de estrutura de população relacionando y a v;

X, S e $Z =$ matrizes de incidência, relacionando y a β, a e u, respectivamente.

As associações foram consideradas significantes quando p<0,001. Além disto, estes p-valores foram submetidos a correção de múltiplos testes, evitando ocorrência de erros do tipo I. O método utilizado para tal correção foi o FDR (False Discovery Rate) (BENAJAMINI e HOCHBERG, 1995) no pacote "*qvalue*" (Storey, 2002) do software R (R DEVELOPMENT CORE TIME). Apenas as associações com q-valores<0,1 foram consideradas significativas. Um *manhatan plot* com valores de –Log10(p-valor) para cada marcador SNP e a respectiva posição nos cromossomos foi também construído pelo software TASSEL.

5.3 Resultados e Discussão

O padrão do decaimento do DL nos 20 grupos de ligação da soja pode ser observado na Figura 1, onde estão plotados os valores significativos de r^2 ($p<0,01$) versus a distância genética em pares de base (pb). Observa-se em todos os grupos de ligação um lento decaímento do DL com o aumento da distância genética, com a presença de grandes blocos em desequilíbrio de ligação nos diferentes cromossomos ao longo de todo o genoma. Sendo assim há pouca necessidade do aumento do número de marcadores para a realização de um mapeamento associativo neste painel de acessos avaliados neste estudo, visto que a resolução do mapa já será boa (FLINT-GARCIA et al., 2003).

Apesar de trabalhos envolvendo acessos do banco de germoplasma de soja informações sobre os padrões de desequilíbrio de ligação em acessos exóticos desta cultura são poucas. Hyten et al. (2007), investigaram o DL em quatro populações distintas de soja: 26 acessos oriundos de *Glycine soja*, 17 variedades crioulas asiáticas de *Glycine max*, 17 ancestrais asiáticos do germoplasma americano e 25 cultivares elite da América do Norte. A extensão do DL encontrado foi maior nos três grupos de *G. max*, segundo os autores isto é devido à correlação entre o desequilibrio, a domesticação e os auto níveis de endogamia presentes nestas populações. Além disto, os autores identificaram alta variabilidade no DL entre as diferentes populações e regiões do genoma analisadas. Esta variabilidade entre diferentes regiões do genoma também pode ser observada neste estudo na Figura 1. Alguns grupos de ligação tais como 5, 9, 11, 12, 13, 16 e 17 apresentam menor extenção de DL quando comparados aos demais. Tal variabilidade pode ser um complicador nas analises de associação.

Os resultados da análise de associação para produtividade de grãos utilizando modelos mistos (*Q+K*) para o modelo utilizando médias fenotípicas dos cinco ambientes corrigidas com os valores de BLUP podem ser observados na Tabela 2. Foram detectadas seis associações, considerando um p-valor<0,0001. Estas associações de marcadores no modelo MLM para a média dos cinco ambientes podem ser observadas no gráfico denominado *Manhattan plot* (Figura 2a). Com aplicação da

correção dos p-valores para múltiplos testes com valores de significância FDR <0,1 verificou-se uma redução de seis para quatro associações (Tabela 2).

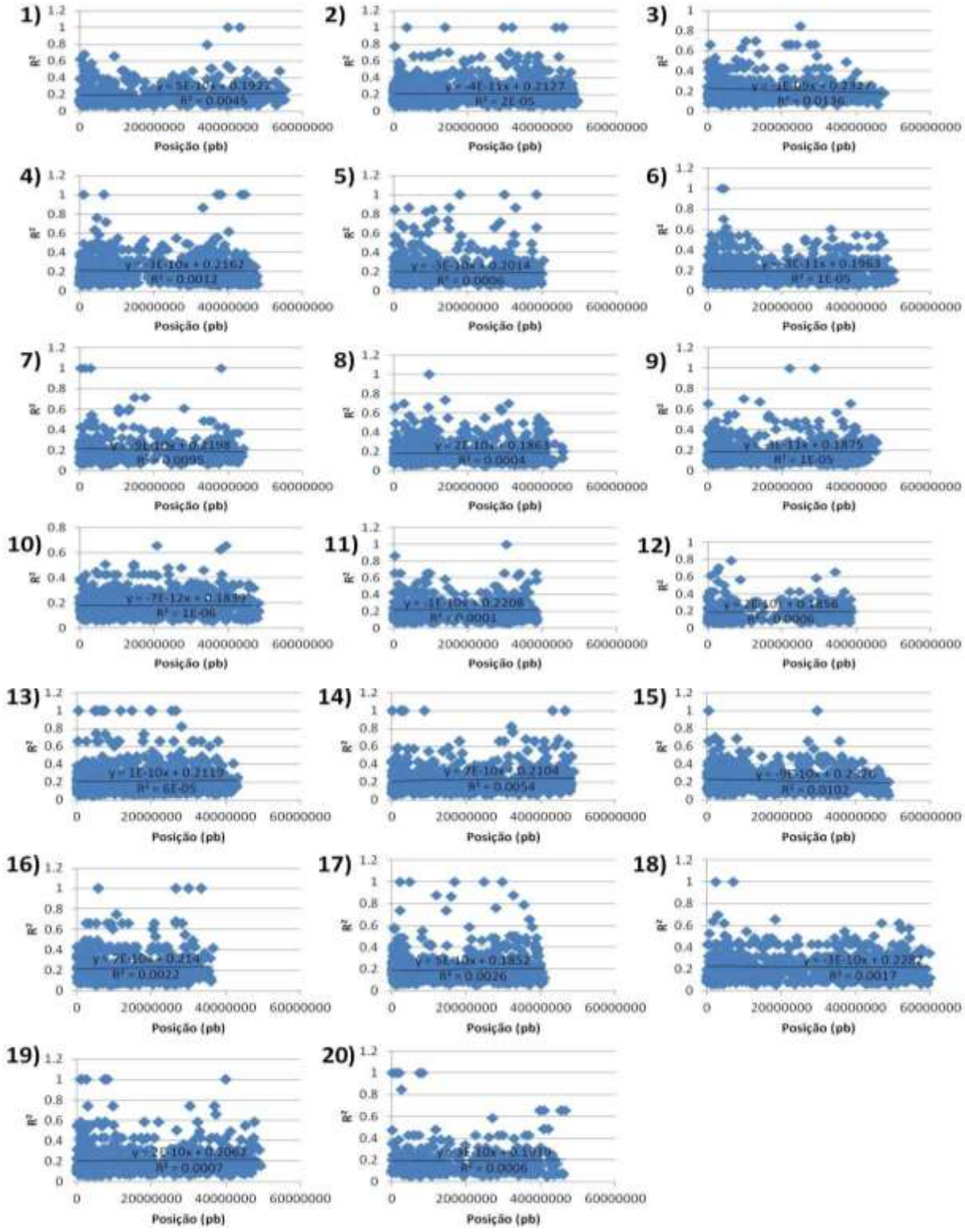

Figura 1 - Decaimento do desequilíbrio de ligação entre pares de marcadores nos 20 grupos de ligação da soja

Tabela 2. Lista dos SNPs associados com a produtividade de grãos em soja utilizando modelos mistos (MLM) para análise conjunta dos 5 ambientes*

Marcador	**Cromossomo**	**R^2**	**p-valor**
AX-90334751	2	0.1841	$6,96E^{-05}$;
AX-90365780	7	0.2348	$4,97E^{-05}$;
AX-90321882	10	0.2348	$4,97E^{-05}$;
AX-90387106	12	0.2348	$4,97E^{-05}$;

*Os 5 ambientes avaliados são: ambiente 1 Piracicaba safra 2012/2013, ambiente 2 Jaboticabal safra 2012/2013, ambiente 3 Ponta Grossa safra 2012/2013, ambiente 4 Piracicaba safra 2013/2014 e ambiente 5 Jaboticabal safra 2013/2014;

Para o modelo considerando as médias ajustadas em cada um dos cinco ambientes e p-valor <0,001 foram encontradas quatro associações para o ambiente um, 11 para o ambiente dois, quatro nos ambientes quatro e cinco. Entretanto, após a correção para múltiplos testes (FDR <0,1), apenas sete associações para o ambiente 2 foram significativas (Tabela 3). Dentre estas apenas três são exclusivas deste local. Tais associações de marcadores podem ser observadas no *Manhattan plot* (Figura 2b).

Dos sete marcadores significativamente associados (Tabela 2 e 3), dois estão localizados no cromossomo dois, grupo de ligação (GL) D1b, e os demais nos cromossomos três, sete, 10, 11 e 12, com respectivos GL, N, M, O, B1 e H. Para estes mesmos marcadores, a porcentagem de variação explicada por cada um, no modelo MLM medida pela estatística R^2, variou de 16% a 23%, indicando que apesar de poucas, as associações que foram detectadas são de grande efeito.

Tabela 3 - Lista dos SNPs associados com a produtividade de grãos em soja utilizando modelos mistos (MLM) para análise do ambiente 2*

Marcador	Cromossomo	R^2	p-valor
AX-90334751	2	0.1629	$1,71E^{-04}$
AX-90365780	7	0.2397	$4,16E^{-05}$
AX-90321882	10	0.2397	$4,16E^{-05}$
AX-90387106	12	0.2397	$4,16E^{-05}$
AX-90362698	2	0.1923	$2,58E^{-04}$
AX-90364328	3	0.2048	$1,58E^{-04}$
AX-90488842	11	0.2048	$1,58E^{-04}$

* Médias ajustadas para o ambiente 2: Jabotical safra 2012/2013

Nas tabelas 4 e 5 observam-se as estimativas dos efeitos alélicos dos SNPs significativos para análise de associação nos dois tipos de abordagem utilizados, na primeira utilizando as médias de BLUP para os cinco ambientes (Tabela 4) e as médias ajustadas para o local 2 (Tabela 5). Para o marcador AX- 90334751 no cromossomo dois (Tabela 3) a diferença para os dois homozigotos, CC e TT para a característica produtividade de grãos é de 888,97 kg ha^{-1}.

O mapeamento associativo (GWAS) tem sido utilizado largamente durante os últimos anos, principalmente devido às limitações relacionadas ao mapeamento convencional, a evolução da genotipagem em larga escala e dos recursos computacionais (KULWAL et al., 2012). Como resultado, vários estudos envolvendo análise de associação em soja foram realizados nos últimos anos (HWANG et al., 2013; MAMIDI et al., 2014; ZHANG et al., 2015).

Os resultados do mapeamento associativo podem ser influenciados pela sub-estruturação da população, acarretada por fatores evolucionários como deriva genética, mutação, seleção e gargalos genéticos. Para evitar a presença de tais fatores na análise de associação e a ocorrência de falsos positivos, os modelos mistos surgiram como uma ótima ferramenta na correção da estrutura de populações (KORTE et al., 2012).

Desenvolvido por Yu et al. (2006) o MLM incorpora duas matrizes, uma de estrutura da população a matriz Q, e a outra matriz de parentesco K, ambas são utilizadas para o controle de associações espúrias no mapeamento associativo. Neste

estudo foi utilizado o emprego de modelos mistos, sendo a matriz Q calculada pelo software STRUCTURE com um número de clusters ou populações k=2. Já a matriz de parentesco K foi calculada pelo método IBD pelo software TASSEL.

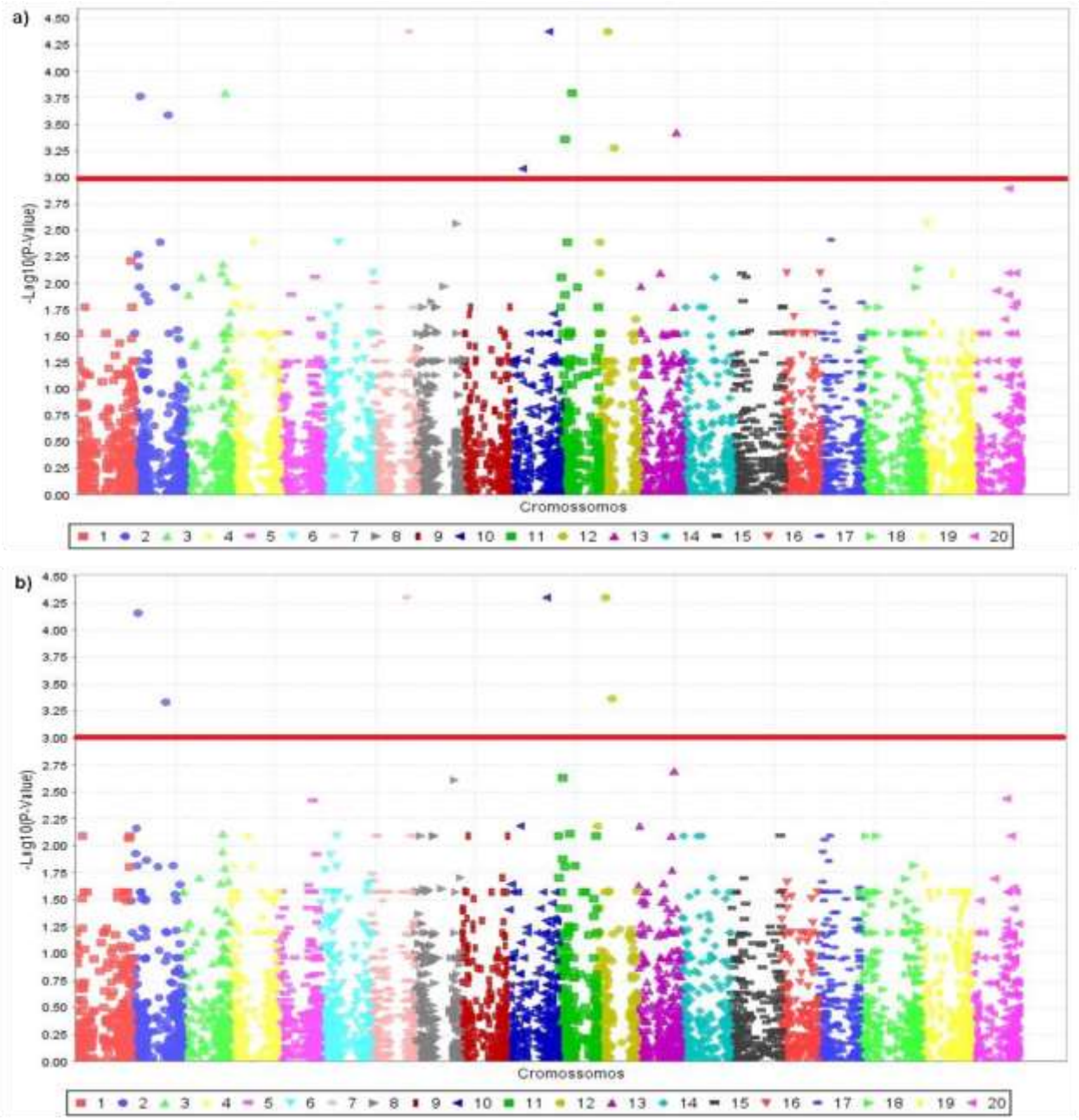

Figura 2. Manhattan plot de valores de –Log10(p-valor) para cada marcador SNP e a respectiva posição nos cromossomos, a linha vermelha representa o nível de significância da associação (−Log 10 p-valor ≥ 3.00, p-valor ≤ 0.001); (a) Modelo MLM para médias ajustadas do ambiente dois; (b) Modelo MLM para médias de BLUP dos cinco ambientes

Finalmente, além da estrutura populacional outro importante fator influenciando a análise de associação é a escolha da população. Fatores como seleção do germoplasma e tamanho da população devem ser levados em conta na seleção da população a ser estudada. Germoplasma exóticos e diversos tendem a minimizar a extensão do desequilíbrio de ligação, pois sofreram várias gerações de recombinação, o que tende a diminuir o DL presente, restando somente o desequilíbrio entre aqueles caracteres e marcadores fortemente ligados (KULWAL et al., 2012).

Tabela 4. Efeitos alélicos dos marcadores SNPs significativos utilizando o modelo MLM com médias BLUP para os 5 ambientes

Marcador	**Loco**	**Alelos***	**Efeito**
AX-90334751	2	C	888.97871
AX-90334751	2	T	-
AX-90365780	7	G	1718.9803
AX-90365780	7	T	630.84613
AX-90365780	7	K	-
AX-90321882	10	C	630.84613
AX-90321882	10	T	1718.9803
AX-90321882	10	Y	-
AX-90387106	12	C	630.84613
AX-90387106	12	T	1718.9803
AX-90387106	12	Y	-

* Código utilizado para os nucleotídeos, derivado de IUPAC: C = C/C; T = T/T; G = G/G; K= G/T; Y= C/T

Neste trabalho foi utilizado um painel contendo 95 genótipos, dentre eles 80 materiais exóticos de diversas partes do mundo e 15 testemunhas comerciais brasileiras. Em estudo realizado por MULATO et al. (2010), os autores detectaram via marcadores microssatélites (SSR) e regiões expressas (EST-SSR), grande diversidade genética nestes mesmos 80 acessos exóticos, indicando a presença de altos níveis de desequilíbrio de ligação. Segundo Nordborg et al. (2002) o decaimento do DL em plantas autógamas tende a ser mais lento do que nas alógamas devido a baixa

recombinação e alta taxa de homozigotos, resultando em altos níveis de DL reportados na literatura (HYTEN et al., 2006).

Outro fator de grande importância que pode influenciar a veracidade dos resultados do mapeamento associativo, assim como no mapeamento de ligação convencional é a interação QTL x ambiente. Nestes estudos é comum a identificação de vários QTLs, entretanto, apenas poucos são detectados quando se considera vários ambientes simultaneamente. Dhanapal et al. (2015) utilizando germoplasma diverso de soja realizou analise de associação para taxa de isótopos de carbono utilizando marcadores SNPs em vários ambientes. Os autores observaram a associação de 39 SNPs em pelo menos dois ambientes e na média dos demais locais.

Tabela 5. Efeitos alélicos dos marcadores SNPs significativos utilizando o modelo MLM com médias ajustadas para o ambiente dois.

Marcador	Loco	Alelos*	Efeito
AX-90334751	2	C	1203.7
AX-90334751	2	T	-
AX-90362698	2	A	1683.3
AX-90362698	2	C	258.37
AX-90362698	2	M	-
AX-90364328	3	C	958.37
AX-90364328	3	T	-597
AX-90364328	3	Y	-
AX-90365780	7	G	2320.7
AX-90365780	7	T	640.22
AX-90365780	7	K	-
AX-90321882	10	C	640.22
AX-90321882	10	T	2320.7
AX-90321882	10	Y	-
AX-90488842	11	A	958.37
AX-90488842	11	G	-597
AX-90488842	11	R	-
AX-90387106	12	C	640.22
AX-90387106	12	T	2320.7
AX-90387106	12	Y	-

* Código utilizado para os nucleotídeos, derivado de IUPAC: C = C/C; T = T/T; G = G/G; K= G/T; Y= C/T; A= A/A; R= A/G; M= A/C

Considerando esta interação QTLs por ambientes, neste estudo envolvendo a análise em cinco ambientes foram utilizadas as médias corrigidas para valores de BLUP nos cinco ambientes avaliados. Foram detectadas quatro associações entre marcador característica (Tabela 2), tais associações também foram detectadas no ambiente dois quando consideradas as médias ajustadas para cada local individualmente (Tabela 3). Com exceção do ambiente dois, em que foram detectadas sete associações, nos demais não foi possível a detecção de marcadores significativos após a correção de FDR, tal fato pode ser atribuído a utilização de poucos indivíduos no mapeamento associativo deste estudo.

A fim de confirmar a veracidade das associações encontradas, os marcadores associados a produtividade de grãos (Tabela 2 e 3) foram confrontados aos QTLs presentes na base de dados de soja *SoyBase* (http://www.soybase.org/). Esta base de dados reúne informações tais como o mapa integrado da soja, busca de sequências e genes previamente publicados.

Para as sete associações detectadas tanto no modelo com as médias da análise conjunta assim como para o ambiente 2, seis foram previamente descritas no *SoyBase*.

Os marcadores AX-90387106 e AX-90334751 foram detectados na região do mesmo QTL, descrito por Orf et al. (1999a). Estes autores, descreveram-no como tendo grande efeito (R^2>10%). O marcador AX-90362698 por sua vez foi detectado na região do QTL descrito previamente por Du et al. (2009), o mesmo ocorreu para o marcador AX-90365780 (Zhang et al., 2004)

Já para as associações exclusivas encontradas no ambiente dois, um marcador foi associado a QTL descrito anteriormente, AX-90364328 (KIM et al., 2012). Este QTL, segundo os autores, explicou 30% da variação para produtividade de grãos, sendo um dos parentais para formação da população F_2 mapeada a PI 90566-1. O marcador AX-90321882, foi associado ao QTL descrito por Rossi et al (2013), com um efeito de R^2 7,9. Entretanto, para a associação detectada no marcador AX-90488842 não houve descrição de QTL em estudos anteriores, indicando uma nova possível região do genoma que possa estar associada à produtividade de grãos. Contudo, esta informação deve ser tratada com cautela visto que esta associação só foi detectada no ambiente 2, sendo assim, há possivelmente forte interação QTL x ambiente.

Sigrist (2010), realizou mapeamento associativo utilizando o mesmo painel de acessos exóticos deste trabalho. Genotipando os acessos com marcadores SSR e fenotipagem em apenas um local o autor encontrou 18 associações para a característica produtividade de grãos. Todas estas associações foram descritas previamente na literatura. Entretanto nenhuma das associações detectadas neste trabalho foi encontrada por Sigrist (2010).

5.4 Conclusões

Foram detectadas sete associações para produtividade de grãos utilizando o modelo MLM. Dentre estas, quatro associações foram observadas utilizando o modelo considerando as médias ajustadas para os cinco ambientes, e também para o ambiente dois quando empregado o modelo considerando as médias de cada ambiente individualmente. As três associações restantes encontradas, foram exclusivas para o ambiente dois, sendo este o único local dentre os cinco avaliados a apresentar associações significativas entre marcador característica.

Com exceção do marcador AX-90488842 todas as outras seis marcas estavam em regiões de QTL previamente reportados na literatura.

Referências

ALLARD, R. W. **Principles of plant breeding**. 2. ed. New York: John Wiley & Sons, 1999. 254 p.

BEDDINGTON, J.. Food security: contributions from science to a new and greener revolution. **Philosophical Transactions Of The Royal Society Of London**, Londres, v. 365, p.61-71, 2010. (Series B).

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. **Journal Of The Royal Statistical Society**, Londres, v. 57, p.289-300, 1995.

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**, Oxford, v. 23, n. 19, p.2633-2635, 2007.

DHANAPAL A. P.; RAY, J. D.; SINGH, S. K.; HOYOS-VILLEGAS, V.; SMITH, J. R.; PURCELL, L. C.; KING, C. A.; CREGAN, P.; SONG, Q.; FRITSCHI, F. B. Genome-wide association study (GWAS) of carbon isotope ratio (δ13C) in diverse soybean [*Glycine max* (L.) Merr.] genotypes. **Theoretical And Applied Genetics**, Berlin, v. 128, n. 1, p.73-91, 2014.

DU, W.; YU, D.; FU, S. Detection of Quantitative Trait Loci for Yield and Drought Tolerance Traits in Soybean Using a Recombinant Inbred Line Population. **Journal Of Integrative Plant Biology**, Beijing, v. 51, n. 9, p.868-878, 2009.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software structure: a simulation study. **Molecular Ecology**, Oxford, v. 14, n. 8, p.2611-2620, 2005.

FLINT-GARCIA, S. A.; THORNSBERRY, J. M.; e BUCKLER, E. S. Structure of linkage disequilibrium in plants. **Annual Review Of Plant Biology**, Palo Alto, v. 54, n. 1, p.357-374, jun. 2003.

GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. **Plant Molecular Biology**, Dordrecht, v. 57, n. 4, p.461-485, 2005.

HYTEN, D. L.; CHOI, I.-Y.; SONG, Q.; SHOEMAKER, R. C.; NELSON, R. L.; COSTA, J. M.; CREGAN, P. B. Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations. **Genetics**, Austin, v. 175, n. 4, p.1937-1944, 2007.

JAKOBSSON, M.; ROSENBERG, N. A.. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics**, Oxford, v. 23, n. 14, p.1801-1806, 2007.

KIM, M. Y.; VAN, K.; KANG, Y. J.; KIM, K. H.; LEE, S. H. Tracing soybean domestication history: From nucleotide to genome. **Breeding Science**, Tokyo, v. 61, n. 5, p.445-452, 2012.

KORTE, A.; VILHJÁLMSSON, B. J.; SEGURA, V.; PLATT, A.; LONG, Q.; NORDBORG, M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. **Nature Genetics**, Nova York, v. 44, n. 9, p.1066-1071, 2012.

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**, San Francisco, v. 9, n. 9, p.e107469, 2014.

MULATO, B. M.; MÖLLER, M.; ZUCCHI, M. I.; QUECINI, V.; PINHEIRO, J. B. Genetic diversity in soybean germplasm identified by SSR and EST-SSR markers. **Pesquisa Agropecuária Brasileira**., Brasília, v. 45, n. 3, p. 276-283, 2010.

NORDBORG, M.; BOREVITZ, J. O.; BERGELSON, J.; BERRY, C. C.; CHORY, J. The extent of linkage disequilibrium in *Arabidopsis thaliana*. **Nature Genetics**, Nova York, v. 30, n. 2, p.190-193, 2002.

ORF, J. H.; CHASE, K.; ADLER, F. R.; MANSUR, L. M.; LARK, K. G. Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. **Crop Science Society of America**, Madison, v.39, p.1652-1657, 1999a.

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

RAMALHO, M.A.P.; SANTOS, J.B. dos; PINTO, C.A.B.P. **Genética na agropecuária**. 4.ed. Lavras: UFLA, 2008. 461p.

RESENDE, M. D. V.; SILVA, F. F. E. ; Resende Jr., M.F.R. . Seleção Genômica Ampla (GWS). In: BORÉM, A.; FRITSCHE-NETO, R. (Org.). **Biotecnologia Aplicada ao Melhoramento de Plantas**. 1ed.Visconde do Rio Branco: Suprema, 2013, v. 1, p. 151-188.

RESENDE, M. D. V.; OLIVEIRA, E. B.; MELINSKI, L. C.; GOULART JUNIOR, F. S.; OAIDA, G. R. P. **Seleção genética computadorizada: SELEGEN "best prediction": manual do usuário**. Colombo: Embrapa- CNPF. 1994. 31 p.

ROSSI, M. Eugenia; ORF, James H.; LIU, Li-jun. Genetic basis of soybean adaptation to North American vs. Asian mega-environments in two independent populations from Canadian × Chinese crosses. **Theoretical And Applied Genetics**, Berlin, v. 126, n. 7, p.1809-1823, 2013.

SIGRIST, M. S. **Mapeamento associativo de locos relacionados à produtividade de grãos em soja**. 81p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2012.

STOREY, J. D. A direct approach to false discovery rates. **Journal Of The Royal Statistical Society: Series B (Statistical Methodology)**, Londres, v. 64, n. 3, p.479-498, 2002.

ZHANG, J.; SONG, Q.; CREGAN, P. B.; NELSON, R. L.; WANG, X.; WU, J.; JIANG, G. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. **Bmc Genomics**, Londres, v. 16, n. 1, p.1-11, 20 mar. 2015.

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

YU, J.M.; PRESSOIR, G.; BRIGGS, W.H.; BI, I. V.; YAMASAKI, M.; DOEBLEY, J. F.; MCMULLEN, M. D.; GAUT, B. S.; NIELSEN, D. M.; HOLLAND, J. B.; KRESOVICH, S.; BUCKLER, E. S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, Nova York, v. 38, n. 2, p.203-208, 2005.

ANEXOS

Anexo A - Lista dos 95 materiais do painel de soja utilizado no mapeamento associativo com respectivo país de origem e província.

(continua)

PI	País	Provincia	PI	País	Provincia
36906	Manchúria	Liaoning	341254	Sudão	-
79861	China	Heilongjiang	341264	Libéria	-
84910	Coréia do Norte	Pyongyang	360851	Japão	-
90251	Coréia do Sul	Seoul	377573	China	-
133226	Indonésia	Java	381660	Uganda	-
145079	Zimbábue	-	381680	Uganda	-
148259	Indonésia	-	398493	Coréia do Sul	Kangwon
148260	África do Sul	North West	407744	China	Jiangsu
153681	El Salvador	-	407764	China	Guangdong
159097	África do Sul	-	416828	Japão	Kanto and Tosan
159927	Peru	Lima	417563	Vietnã	-
164885	Guatemala	Escuintla	417581	EUA	-
165524	Índia	-	417582	EUA	-
166141	Nepal	Khatmandu	427276	China	Guangdong
170889	África do Sul	Transvaal	438301	Coréia do Norte	-
171437	China	Sichuan	90577	China	Liaoning
172902	Turquia	Artvin	159922	Peru	Lima
189402	Guatemala	Sacatepequez	209839	Nepal	Gandaki
200832	Mianmar	Kachin	222546	Argentina	Buenos Aires
203400	França	-	240665	Filipinas	Luzon
203404	Japão	-	281898	Malásia	-
204333	Suriname	-	281911	Filipinas	-
204340	Suriname	-	284816	Malásia	-
205384	Paquistão	-	306712	Tanzânia	Tanga
205912	Tailândia	-	281907	Malásia	-
210178	Taiwan	-	IAC100	Brasil	Campinas
210352	Moçambique	Maputo	Paranagoiana	Brasil	-
212606	Afeganistão	Nangarhar	A7002	Brasil	Mato Grosso
215692	Israel	Central	CD215	Brasil	Mato Grosso do Sul
222397	Paquistão	Northern Areas	Conquista	Brasil	Mato Grosso
222550	Argentina	Buenos Aires	Pintado	Brasil	Mato Grosso
229358	Japão	Kanto district	Sambaíba	Brasil	-
239237	Tailândia	-	Dowling	EUA	Texas
253664	China	Shanghai	Shira Nuhi	Japão	Shikoku
259540	Nigéria	-	Kinoshita	Japão	Shikoku
265491	Peru	Lima	Orba	Indonésia	-

Anexo A - Lista dos 95 materiais do painel de soja utilizado no mapeamento associativo com respectivo país de origem e província.

(conclusão)

PI	País	Provincia		PI	País	Provincia
265497	Colômbia	Cundinamarca		Bing Nan	China	-
274454-A	Japão	-		227687	Japão	Okinawa
274454-B	Japão	-		171451	Japão	Kanagawa
274507	China	-		VMáx	Brasil	-
283327	Taiwan	-		Potência	Brasil	-
285095	Venezuela	Aragua		JAB 00-05-6/763D	Brasil	Jaboticabal
297550	Rússia	-		JAB 00-02-2/2J3D	Brasil	Jaboticabal
306702	Tanzânia	-		LQ 1050	Brasil	Piracicaba
315701	EUA	-		LQ 1505	Brasil	Piracicaba
322695	Angola	-		LQ 1421	Brasil	Piracicaba
331793	Vietnã	Saigon		LQ 1413	Brasil	Piracicaba
331795	Vietnã	Saigon				

Anexo B - Análise de Deviance (ANADEV) conjunta de 5 ambientes para 9 características em 95 linhagens de soja

Efeitos	APM	PEG	VA	AC	MCS	NDM	APF	IPV	OLEO
Genótipos	9874,29**	6306,73**	740,09**	1256,92**	4481,89**	6124,61**	3635,94**	3218,6**	1247,12**
Blocos	9371,51^{ns}	6265,81^{ns}	644,34^{ns}	1035,19**	4166,84^{ns}	6047,67**	3575,42^{ns}	3151,42^{ns}	1188,56^{ns}
Genótiposxlocais	9511,34	6293,97**	813,87**	1106,05**	4295,11**	6071,66**	3683,74**	3263,45**	1219,99**
Modelo Completo	9369,64	6263,6	644,24	1028,41	4166,67	6036,21	3575,34	3150,5	1187,66
CV%	14,03	21,26	22,99	28,63	16,4	7,78	13,88	22,95	7,18
Média	95,99	40,29	2,61	2,51	13,67	135,81	70,64	12,75	18,06

*Qui – quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5% e 1%, respectivamente