Universidade Estadual de Londrina

ANDRÉ LUIZ DE LIMA PASSIANOTTO

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA DA SOJA E SEU USO NO MAPEAMENTO ASSOCIATIVO DE CARACTERÍSTICAS SIMPLES E COMPLEXAS

Londrina
2014

Universidade Estadual de Londrina

Brasileira de Pesquisa Agropecuária

ANDRÉ LUIZ DE LIMA PASSIANOTTO

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA DA SOJA E SEU USO NO MAPEAMENTO ASSOCIATIVO DE CARACTERÍSTICAS SIMPLES E COMPLEXAS

Londrina
2014

ANDRÉ LUIZ DE LIMA PASSIANOTTO

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA DA SOJA E SEU USO NO MAPEAMENTO ASSOCIATIVO DE CARACTERÍSTICAS SIMPLES E COMPLEXAS

Tese apresentada ao Programa de Pós–Graduação, em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de doutor.

Orientador: Prof. Dr. Ricardo Vilela Abdelnoor

Londrina
2014

Catalogação elaborada pela Divisão de Processos Técnicos da Biblioteca Central da Universidade Estadual de Londrina.

Dados Internacionais de Catalogação-na-Publicação (CIP)

P288i Passianotto, André Luiz de Lima.
Identificação de polimorfismos de nucleotídeo único no genoma da soja e seu uso no mapeamento associativo de características simples e complexas / André Luiz de Lima Passianotto. – Londrina, 2014.
69 f. : il.

Orientador: Ricardo Vilela Abdelnoor.
Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2014.
Inclui bibliografia.

1. Soja – Melhoramento genético – Teses. 2. Sequência de nucleotídeos – Teses. 3. Polimorfismo (Genética) – Teses. 4. Nematoda em plantas – Teses. 5. Marcadores biológicos – Teses. I. Abdelnoor, Ricardo Vilela. II. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. III. EMBRAPA. IV. Título.

CDU 631.52:633.34

ANDRÉ LUIZ DE LIMA PASSIANOTTO

IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA DA SOJA E SEU USO NO MAPEAMENTO ASSOCIATIVO DE CARACTERÍSTICAS SIMPLES E COMPLEXAS

Tese apresentada ao Programa de Pós–Graduação, em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito parcial para a obtenção do título de doutor.

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Ricardo Vilela Abdelnoor
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA

_____

Prof. Dr. François Belzile
Université Laval

_____

Prof. Dr. Ney Sussumu Sakiyama
Universidade Federal de Viçosa – UFV

_____

Prof. Dr. Laurival Antonio Vilas-Boas
Universidade Estadual de Londrina – UEL

_____

Prof. Dr. Francismar Correa Marcelino-Guimarães
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA/Soja

Londrina, 19 de Fevereiro de 2014.

Aos meus pais, minha irmã e à minha noiva, primeiro por fazerem parte da minha vida e segundo pelo constante incentivo principalmente nos momentos onde parti em busca de conhecimento.

Dedico.

“E de repente sua vida começa a ganhar um novo sentido, precisamos aprender a superar os desafios que nos são apresentados, só assim alçaremos vôos mais altos.”

AGRADECIMENTOS

A Deus pela maior graça dada a uma pessoa, a oportunidade de viver a vida.

À Universidade Estadual de Londrina e ao Programa de Pós-Graduação em Genética e Biologia Molecular, pela oportunidade de realizar o curso de Mestrado.

À CAPES e CNPq pela concessão de bolsa de estudos.

À Embrapa pela disponibilização de estrutura física para a realização deste trabalho.

Ao professor doutor Ricardo Vilela Abdelnoor, pela ajuda no andamento do trabalho, correções e pelo incentivo a minha pessoa. Muito Obrigado.

Ao professor doutor François Belzile, por me auxiliar nas analises de bioinformática, na discussão dos dados, bem como no aprendizado compartilhado com a minha pessoa. Muito Obrigado

Aos professores do Programa de Pós-Graduação em Genética e Biologia Molecular obrigado pelas discussões, conselhos e incentivo.

Aos pesquisadores da Embrapa, doutor Álvaro M. R. Almeida, doutor Eliseu Bineck, doutor Carlos Alberto Arrabal Arias e doutor Marcelo Fernandes de Oliveira, pelo apoio técnico-científico e pelo incentivo.

Aos estagiários e bolsistas que fazem e fizeram parte do Laboratório de Biotecnologia Vegetal da Embrapa-Soja, obrigado pelo convívio, alegrias, tristezas e aprendizado adquirido nesses anos.

Aos Técnicos do Laboratório, Silvana Marin, César Silveira, Nilson Vieira, Vera Pieronte e Márcia Kuwahara, Danielle Cristina Gregório e Renan Novaes Milagres pela imensa ajuda que cada um deles disponibilizou para a realização deste trabalho. Muito Obrigado.

À secretaria do Programa de Pós-Graduação em Genética e biologia Molecular na pessoa da secretária Sueli e da coordenadora Ana Lúcia, pela constante disponibilidade desprendida a minha pessoa, obrigado pelo constante auxílio. Muito Obrigado a vocês pela ajuda.

Ao pessoal do IBIS, Maxime, Aurelie, Huma, Elmer, Martine, Sebastien, Remi, Romain, Olfa, Gabrielle, Suzane, Érika, Renato, Valérie eu não possuo

palavras para agradecer à vocês por tornar minha estadia muito mais interessante e divertida. Vou levar cada um de vocês no coração.

Aos meus pais tenho que agradece-los por tudo isso, vocês também são os merecedores desse título, a maior conhecimento aplicado neste estudo foi adquirido de vocês. Obrigado por serem meus pais. Tenho muito orgulho de vocês dois e junto com a Iza tenho satisfação imensa de dizer que vocês são a minha família e participaram em conjunto dessa conquista.

À minha noiva Thayse Maria Marestoni, pela imensa ajuda nas discussões sobre o trabalho, idéias, revisões, compreensão e carinho disponibilizado a mim, mesmo nos momentos mais difíceis da realização das atividades. Sem sua companhia, tudo teria sido muito mais difícil.

Enfim, a todos que de alguma forma contribuíram para a realização deste trabalho, meus mais sinceros agradecimentos.

PASSIANOTTO, André Luiz de Lima. **Identificação de polimorfismos de nucleotídeo único no genoma da soja e seu uso no mapeamento associativo de características simples e complexas**. 2014. 69f. Doutor em Genética e Biologia Molecular - Universidade Estadual de Londrina, Londrina, 2014.

RESUMO

A soja *Glycine max* (L.) Merrill, é uma oleaginosa com grande destaque na economia mundial. O Brasil é o segundo maior produtor de soja, e na safra 2012/2013 apresentou produtividade média de 2.933 kg/ha, totalizando 81 milhões de toneladas, contribuindo para o desenvolvimento de inúmeras regiões e sendo um dos pilares chefes do agronegócio brasileiro. O sucesso desta cultura se baseia no seu progresso genético, adaptação ao ambiente brasileiro e melhoria nas técnicas de cultivo. Cultivares foram melhoradas ao longo dos anos com o intuito de se obter materiais aclimatados as fronteiras agrícolas brasileiras e com melhor produtividade. As doenças são um dos fatores que acometem a cultura e impactam diretamente na produtividade da soja brasileira caso com o do nematóide *Meloidogyne incógnita* o qual acomete lavouras de soja e seu parasitismo pode causar severos danos à lavoura, variando de acordo com o grau de infestação. Recentemente, com as técnicas de sequenciamento de nova geração, novas metodologias visando a rápida identificação de polimorfismos e seu uso nos estudos de mapeamento associativo surgiram. Entre elas destaca-se a Genotipagem por Sequenciamento (GBS). Este trabalho teve por objetivo a identificação de polimorfismos de nucleotídeo único (SNP) e a validação da tecnologia de GBS para mapeamento de características simples e complexas em soja. Utilizando uma população de 165 cultivares brasileiros, características como cor da pubescência, cor da flor e tolerância a glifosato foram mapeadas nos cromossos 6, 13 e 2 respectivamente corroborando com os estudos prévios de mapeamento apresentados por estas características. Além disso, utilizando 194 genótipos de soja resistentes e suscetíveis ao nematóide *Meloidogyne incógnita*, foi possível a identificação de 17.530 SNPs e o mapeamento de um QTL para resistência ao nematoide de galha, no cromossomo 10.

Palavras-chave: Melhoramento genético. Seqüência de nucleotídeos. Genética vegetal.

PASSIANOTTO, André Luiz de Lima. **Identification of single nucleotide polymorphisms on the soybean genome and its use in association mapping for single and complex traits**. 2014. 69p. Thesis (Phd in Genetics and Molecular Biology) - Universidade Estadual de Londrina, Londrina, 2014.

ABSTRACT

Soybean *Glycine max* (L.) Merrill, is an oilseed crop with great prominence in the world economy. Brazil, the world second largest soybean producer, reached a yield of 2933 kg/ha for the 2012/2013 season and totaling 81 million tonnes, contributing to the development of several regions and becoming one of the most important commodities of Brazilian agribusiness. The success of this crop is based on genetic advances, environmental adaptation and better crop practices. Soyban cultivars have been improved over the years in order to obtain materials adapted to diferent Brazilian regions frontiers and better yield. However, diseases are one of the main factors affecting the crop and can make a great impact on Brazilian soybean crop yields. The root knot nematode (RKN), *Meloidogyne incognita,* affects soybean yield and the infection can cause severe damage to the crop according to the degree of infestation. Recently, with techniques based on new generation sequencing, new methodologies aiming a fast identification of polymorphisms and its direct use on association mapping became available. This study aimed to identify single nucleotide polymorphisms (SNPs) and validate the GBS methodology for mapping single and complex traits. Based on a population of 165 Brazilian cultivars, the glyphosate tolerance, pubescence color and flower color were correctly mapped on chromosomes 2, 6 and 13, respectively. In addition, 194 soybean accessions were also evaluated by GBS, allowing the identification of 17,530 SNPs and mapping a QTL for RKN resistance, on chromosome 10.

Keywords: Soybean – Breeding. Nucleotide sequence. Plant genetics.

SUMÁRIO

1.1 Introdução

Atualmente, a cultura da soja [*Glycine max* (L.) Merrill] é uma das principais culturas brasileiras, gerando empregos em setores econômicos diversos e contribuindo massivamente para o desenvolvimento do país. A área plantada se espalha por 16 estados da federação onde duas regiões produtoras distintas (Centro-Sul e Norte-Nordeste) compartilham os 27.715 mil ha cultivados na safra de soja 2013/2014, com uma produção estimada de 81 milões de tonelada (CONAB, 2013). Na safra 2012/2013, a produtividade média geral brasileira foi 2.933 kg/ha, onde 25.042 mil ha produziram 81,2 milhões de toneladas do grão, um incremento de 22,4 % comparando a safra anterior (2011/2012) (CONAB, 2013). O complexo soja ocupa lugar de destaque no mercado de *commodities* nacional, gerando sucessivos superávits na balança comercial brasileira. A exportação referente a soja se estende desde o grão até seus derivados e estão focadas principalmente nos grandes mercados consumidores da China e EUA, (CONAB, 2013).

No entanto, o entendimento de suas caraterísticas agronômicas, como tolerância a estresses bióticos e abióticos, é de fundamental importância para que o potencial produtivo da soja seja explorado ao máximo. Assim, com o objetivo de explorar a amplitude de seu genoma e entender melhor suas características, estudos moleculares vem sendo feitos ao longo das ultimas décadas, tendo como base diferentes tipos de marcadores moleculares. Os marcadores RFLP (*restriction fragment length polymorphisms*)(Zhang et al., 2013) foram os primeiros utilizados em estudos de diversidade genética (Keim et al., 1989) e na geração de mapas gênicos (Keim et al., 1990). Posteriormente os marcadores RAPD (*random amplified polymorphic DNA)* se destacaram pela sua facilidade, rapidez e versatilidade se tornando uma ferramenta poderosa para estudos genético-moleculares (Abdelnoor et al., 1995), possibilitando a

rápida identificação de polimorfismos ligados a locos de resistência à doenças, como por exemplo o gene *Co-4* em feijoeiro (Cardoso e Arruda, 1998) e em nematóides em soja (Silva, 2001; Schuster, 2001). A especificidade gerada pelos sítios de restrição do RFLP aliada à praticidade da amplificação por PCR permitiu o desenvolvimento dos marcadores AFLP (*amplified fragment length polymorphism*)(Maughan et al., 1996). Marcadores AFLP foram muito utilizados para estimar coeficientes de similaridade, como um estudo sobre a similaridade genética em 186 cultivares de soja (Bonato, et al., 2006).

Tendo em vista alto grau de polimorfismo e o nível informativo apresentados por sequências repetitivas no genoma, Akkaya et al. (1992) conduziram experimentos em soja, para verificar a freqüência na natureza de polimorfismo e o modo de herança dos marcadores microssatélites (simple sequence repeat) . Desde então a análise de marcadores microssatélites ou SSR markers se mostraram uma poderosa ferramenta no estudo de mapeamento genômico, da genética de populações, do manejo e da conservação de espécies (Schuster et al., 2001). Entre outras aplicações, os marcadores microssatélites têm sido utilizados também na caracterização da diversidade genética, na identificação de germoplasma e no estudo da dinâmica de populações (Rongwen et al., 1995; Senior et al.,1998). Até o momento, um grande número de loci de resistência a doenças em soja tem sido mapeado com o uso de marcadores SSR (Silva et al., 2010; Yamanaka et al., 2011; Yamanaka et al., 2013; Lemos et al., 2011).

Marcadores moleculares têm por objetivo auxiliar programas de melhoramento através da seleção assistida, facilitando assim rápidos ganhos genéticos (Jannink *et al*., 2010). A detecção e a exploração da variação genética sempre foi uma parte fundamental do melhoramento de plantas. Variações presentes no DNA associadas a

caracteres de interesse agronômico foram abordadas durante as duas ultimas décadas com diferentes tipos de marcadores moleculares (Varshney et al., 2009).

Mais recentemente, o aumento do conhecimento do genoma via análises “in silico” permitiu a identificação de variações que representam o tipo mais abundante de variação genética existente, os polimorfismos de base única (Single Nucleotide Polymorphism - SNP). Estes possuem uma ampla aplicabilidade em pesquisas genômicas modernas (Gaur et al., 2012). Acessando as informações genômicas da soja foi possível por exemplo, associar variações alélicas à rendimento e ganhos de produção visando médias maiores de produtividade (Hao, *et al*., 2012). Vários estudos envolvendo a descoberta e uso de SNPs já foram relatados em diferentes espécies de plantas e animais. Em soja, vários estudos já foram realizados com enfoque na identificação de caracteres agronômicos de interesse, como produtividade e resistência a patógenos (Wu et al., 2010, Hyten et al 2010, Hao, et al., 2012; Bastien, et al 2013).

Nos últimos anos, a genômica sofreu uma expansão exponencial e os dados gerados pelas novas tecnologias de sequenciamento (NGS), permitiram que o sistema baseado em SNP se tornasse o tipo de marcador mais utilizado por acadêmicos e melhoristas nos estudos genômicos de características complexas (Deschamps & Campbell, 2010). Devido a sua importância estratégica, a soja foi escolhida como planta modelo dentre as leguminosas, tendo o genoma da cultivar Williams 82 completamente sequenciado (Schmutz *et al*. em 2010), e mais recentemente, alguns esforços de re-sequenciamento de genomas parciais ou completos tem permitido a comparação das sequencias e a identificação de um grande número de SNPs (Sonah et al., 2013).

No entanto, definir a variação genética existente no germoplasma da soja e correlacioná-la com características presentes na planta ainda são os grandes desafios

atuais da cultura (Schmutz *et al*., 2010). A seleção assistida por marcadores (SAM) obteve sucesso com a descoberta e mapeamento de característica simples, controladas por um único gene. Entretanto quando nos referimos a características poligênicas o mesmo sucesso não foi alcançado (Bernardo et al., 2008). Populações biparentais tem sido tradicionalmente utilizadas para a descobertas de QTLs por meio de análises com marcadores moleculares, no entanto, essas populações não representam a diversidade alélica como um todo (Jannink *et al*., 2010). Progênies de tamanhos muito grandes teriam que ser geradas para se obter uma probabilidade razoável de recuperar os genótipos com combinações alélicas favoráveis à maioria dos QTLs (Resende, et al., 2012). O entendimento de características genéticas complexas demandam um conhecimento avançado sobre a estrutura genômica da característica, sua localização bem como a extensão do QTL no cromossomo são informações básicas requeridas para tal. O desenvolvimento de marcadores moleculares de baixo custo, que proporcionem uma cobertura do genoma suficientemente densa para a dissecção de características complexas pode certamente contribuir para o aumento do uso de marcadores nos programas de melhoramento (Gaur, et al., 2012).

Nos últimos anos, com o aumento da quantidade de dados gerados pelas tecnologias de sequenciamento de nova geração (NGS – New Generation Sequencing), houve um significativo impulso na geração de dados de sequenciamento e também nos estudos de associação genômica ampla (GWAS – Genomic Wide Association Studies). Os estudos de GWAS postulam que genótipos com a mesma característica compartilham as mesmas regiões gênicas. Aliando a informação genotípica disponibilizada pelo sequenciamento ao conhecimento prévio do mapeamento é possível a predição de fenótipos específicos, atuando de base para o desenvolvimento de cultivares melhoradas em um menor tempo, com maiores níveis

de resistência ou tolerância a fatores bióticos e/ou abióticos por exemplo (Vashney et al., 2009).

Uma ampla atenção tem sido empregada não só na descoberta de SNPs mas principalmente no desenvolvimento de metodologias de genotipagem que permitam trabalhar com estes polimorfismos (Edwards *et al*., 2008; Ganal *et al*., 2009; Varshney *et al*., 2009). Plataformas de genotipagem onde o conhecimento prévio dos SNPs é requerido, podem ser customizadas de acordo com SNPs pertinentes ao estudo (Hwang, et al., 2014; Mancini, et al 2014). O BeadChip® (SoySNP50K), desenvolvido por Song, et al. (2013), teve como base as sequencias de seis cultivares de soja (*G. max*) e dois acessos de soja selvagem (*G. soja*), permitindo a identificação e validação de cerca de 50 mil SNPs entre estes genótipos. A aplicabilidade desta técnica permitiu a identificação de 40 SNPs associados à teores de óleo e proteína nas sementes, em 17 regiões genômicas (Hwang et al., 2014).

Em uma outra abordagem, a variabilidade genética presente na população estudada não é previamente conhecida e se faz necessário o sequenciamento para que essa possa ser acessada e associada com características de interesse (van Orsouw, et al., 2007; Elshire, et al,.2011, Trebbi, et al. 2011; Sonah, et al., 2013, Bastien, et al,. 2013). Técnicas que envolvem o sequenciamento parcial de genomas através da redução de complexidade tem sido usadas com resultados promissores para estudos de GWAS. A tecnologia CRoPS (redução de complexidade em sequencias polimórficas), onde o emprego da técnica de AFLP produz fragmentos sequenciados via piro-sequenciamento foi descrita por van Orsouw et al (2007) em milho. Trebbi et al (2011) visando a descoberta de SNPs em trigo utilizou a técnica, validando-a para analises genômicas amplas em espécies poliplóides.

Dentre as técnicas de redução de complexidade mais utilizadas, a genotipagem por sequenciamento (GBS – *Genotyping by Sequencing*) (Elshire *et al*., 2011) se destaca pela sua multiplexidade, baixo custo por amostra e alta qualidade dos SNPs gerados. Com essa técnica, 96 amostras podem ser sequenciadas simultaneamente em cada canal da flowcell (Figura 01), os genótipos disposto na placa são individualmente clivados via enzima de restrição *Ape*KI, posteriormente são adicionados ao DNA clivado adaptadores contendo em uma das extremidades o primer de sequenciamento aliado ao barcode de identificação e na outra somente o primer para sequenciamento. Os barcodes possuem a variação de 4 a 8 nucleotídeos e são únicos a cada um dos genótipos, permitindo assim identificar o DNA do individuo dentro do pool de material genético dos 96 genótipos sequenciados. A construção da biblioteca é realizada via PCR e o sequenciamento utiliza a plataforma Hiseq 2500 onde uma média de 1,2 milhões de reads por barcode foi relatada por Passianotto et al., 2014. Os dados gerados pelo sequenciamento atual prove em média 200 milhões de reads por canal, a regiões genicas sequenciadas tendem a eliminar regiões repetitivas e focam em regiões onde os transcritos estão situados. O tratamento dos dados consiste em identificar os reads via barcode e separá-los criando arquivos onde estão contidos todos os reads do genótipo. O alinhamento destes reads é realizada de acordo com a versão do genoma escolhida pelo usuário, posteriormente são identificados os SNPs contidos nos materiais e é criado um catálogo de SNPs. Este processo é totalmente executado pelo pipeline descrito pelo departamento de ciências em plantas e instituto de integração em biologia de sistemas (IBIS) da Universidade Laval (Sonah et al, 2013).

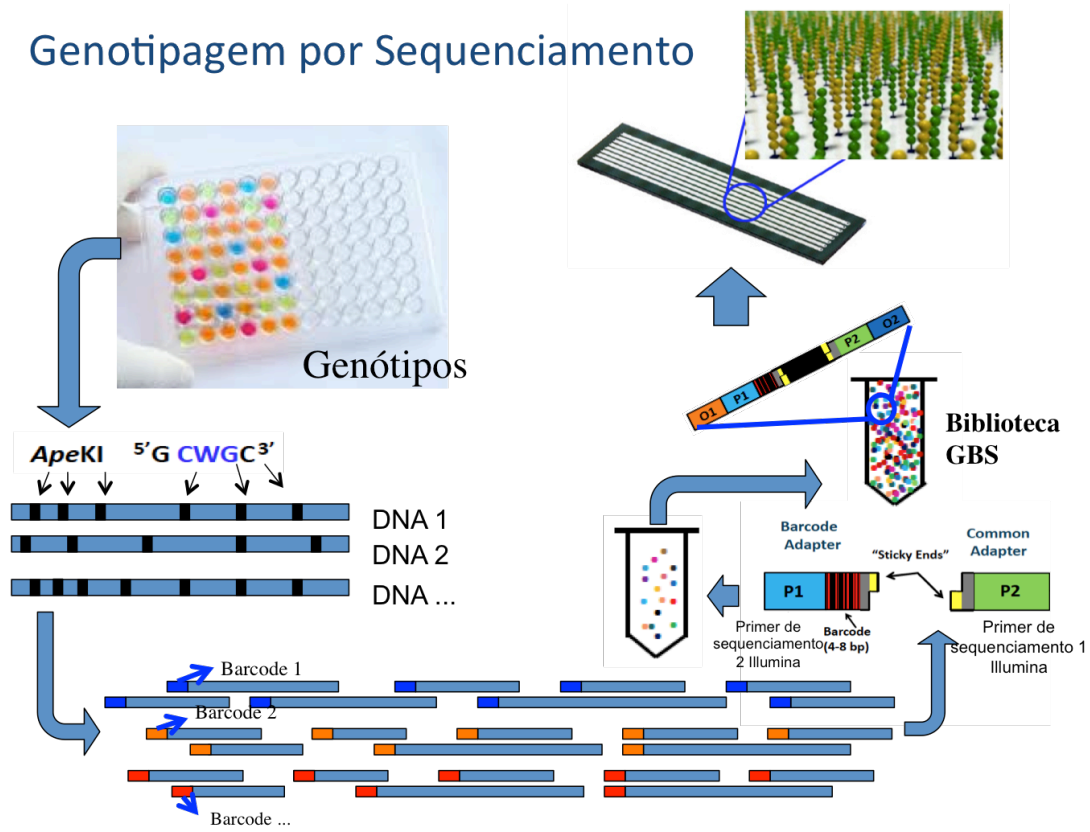

Figura 1 . **Esquema geral da técnica de Genotipagem por Sequenciamento**. A digestão dos DNAs ocorre com a enzima de restrição *Ape*K1, os DNAs digeridos são ligados a barcodes onde são posteriormente acoplados a *primers* de sequenciamento, a amplificação destes proporciona a construção da biblioteca GBS onde é disposta em um dos canais da FLowcell.

Analises genômicas com o uso de GBS se mostraram promissores tanto na geração de dados genotípicos, bem como na construção de modelos genômicos, como no estudo visando seleção genômica em *Triticum aestivum L*, uma cultura com o genoma complexo e poliploide (Poland, et al., 2012). Apesar da maior parte dos estudos realizados ate o momento utilizar a plataforma de sequenciamento HiSeq 2000, essa técnica foi avaliada também com a plataforma Ion Torrent (Life Technologies, Carlsbad, CA), em um estudo com cevada. Nesse estudo, uma taxa máxima de discordância entre os genótipos fornecidos pelas técnicas foi de apenas 2%, sendo que apenas pequenas mudanças nos adaptadores de sequenciamento descritos para a plataforma Hiseq 2000 foram realizadas para sua utilização no Ion Torrent (Mascher, et al., 2013).

Apesar de recente, algumas variações nessa técnica já foram avaliadas. Em um estudo envolvendo a cultura da soja, modificações envolvendo o protocolo de preparação das bibliotecas e dos “pipelines” de identificação de SNPs permitiu um aumento de ao menos 40% no número de SNPs de alta qualidade (Sonah et al., 2013). Contudo, há a necessidade de otimização do GBS para utilização em diferentes espécies, bem como também o desenvolvimento de ferramentas robustas em bioinformática para análise dos dados gerados (Elshire *et al*., 2011).

No contexto de inovação tecnológica aplicada ao melhoramento a primeira parte do presente estudo atua desenvolvendo, descobrindo e validando marcadores SNPs e a segunda parte constitui o mapeamento por associação, onde por meio da análise genotípica e fenotípica dos materiais disponibilizados, promove a detecção de SNPs associados a características de interesse.

1.2 References

Abdelnoor R. V.; Barros E. G.; Moreira M. A. Determination of diversity within Brazilian soybean germplasm using random amplified polymorphic DNA techniques and comparative analysis with pedigree data. Brazilian Journal of Genetics. v. 18 p.1265-273, 1995.

Akkaia, M.S.; Bhagwat. A.A.; Creagan. P.B. Lenght polimorphism of simple sequence repeat DNA in soybean. Genetics. v.132, n.3, p. 1131-1139. 1992.

Bastien, M.; Sonah, H; Belzile. (2013). Genome wide association mapping of Sclerotinia sclerotiorum resistance in soybean with a genotyping by sequencing approach. Plant Genome. doi:10.3835/plantgenome2013.10.0030.

Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. CropSci 2008; 48:1649–64.

Bonato, A.L.V.; Calvo, E.S.; Geraldi, I.O.; Arias, C.A.A. Genetic similarity among soybean (Glycine max (L) Merrill) cultivars released in Brazil using AFLP markers. Genetics and Molecular Biology, v.29, p.692-704, 2006.

Companhia Nacional de Abastecimento – CONAB. Acomp. safra bras. grãos, Quarto Levantamento, Brasília, 1:67. Disponível em: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_01_10_15_07_19_boletim_graos_janeiro_2014.pdf.

Deschamps, S; Campbell, M.A.(2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery.Mol Breed 25:553–570

Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract. Accessed 21 January 2014.

Edwards, D.; Forster, J. W.; Cogan, N. O. I.; Batley, J. ; Chagné, D.. (2007) Single Nucleotide Polymorphism Discovery. In Oraguzie, NC; Rikkerink, EHA; Gardiner, SE; Silva, HN de. Association Mapping in Plants, Springer New York, 278 p.

Ganal M.W.; Altmann T.; Röder M.S. (2009). SNP identification in crop plants. Curr Opin Plant Biol. 12:211-217.

Gaur R, Azam S, Jeena G, Khan AW, Choudhary S, et al. (2012) High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (Cicer arietinum L.). DNA Res 19: 357–373. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3473369&tool=pmcentrez&rendertype=abstract. Accessed 12 February 2014.

Hao D, Cheng H, Yin Z, Cui S, Zhang D, et al. (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. Theor Appl Genet 124: 447–458. Available: http://www.ncbi.nlm.nih.gov/pubmed/21997761. Accessed 2 July 2013.

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9: 166–177. Available:

Keim P, Olson TC, Shoemaker RC (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newslett 15:150– 152

Keim. P.. B.W. Diers. T.C. Olson. and R.C. Shoemaker. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735– 742. 1990.

Hwang E, Song Q, Jia G, Specht JE, Hyten DL, et al. (2014) A genome-wide association study of seed protein and oil content in soybean: 1–12.

Hyten DL, Cannon SB, Song Q, Weeks N, Fickus EW, et al. (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11: 38. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817691&tool=pmcentrez&rendertype=abstract.

Lemos NG, Lucca e Braccini A, Abdelnoor RV, Oliveira MCN, Suenaga K, et al. (2011) Characterization of genes Rpp2, Rpp4, and Rpp5 for resistance to soybean rust. Euphytica 182: 53–64. Available: http://link.springer.com/10.1007/s10681-011-0465-3. Accessed 26 July 2013.

Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. PLoS One 8: e76925. Available: http://www.ncbi.nlm.nih.gov/pubmed/24098570. Accessed 9 October 2013.

Maughan PJ, Maroof MAS, Buss GR. (1996) Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor Appl Genet 93:574–579.

MOREIRA. Maurilio Alves . Identification of a new major QTL associated with resistance to soybean cyst nematode (Heterodera glycines). Theoretical And Applied Genetics. Alemanha. v. 102. p. 91-96. 2001.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, et al. (2012) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome J 5: 103. Available: https://www.crops.org/publications/tpg/abstracts/5/3/103. Accessed 31 October 2013.

Resende MFR Jr, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012b) Accurate of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190:1503–1510

Rongwen J, Akkaya MS, Bhahwat AA, Lavi U, Cregan PB (1995) The use of microsatellite DNA markers for soybean genotype identification. Theor Appl Genet 90: 43-48.

Senior, M. L., J. P. Murphy, M. M. Goodman and C. W. Stuber, 1998 Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci. 38: 1088–1098.

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, et al. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8: e54603. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3553054&tool=pmcentrez&rendertype=abstract. Accessed 21 September 2013.

Song Q, Hyten DL, Jia G, Quigley C V, Fickus EW, et al. (2013) Development and Evaluation of SoySNP50K , a High- Density Genotyping Array for Soybean. 8: 1–12. doi:10.1371/journal.pone.0054985.

Schuster. I. ; Abdelnoor. R. V. ; Marin. S. R. R. ; Carvalho. V. P. ; Kiihl. R. A. S. ; Silva. João Flávio V ; SEdiyama. Carlos S ; Barros. Everaldo Gonçalves de ; Silva. J.F.V.; L.C.B.C. Ferraz; C.A.A. Arias & R.J.E. Abdelnoor. Identificação de marcadores moleculares de microssatélite associados à resistência de genótipos de soja a Meloidogyne javanica. Nematologia Brasileira. 25(1):79-83. 2001.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183. Available: http://www.ncbi.nlm.nih.gov/pubmed/20075913. Accessed 18 September 2013.

Trebbi D, Maccaferri M, de Heer P, Sørensen A, Giuliani S, et al. (2011) High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theor Appl Genet 123: 555–569. Available: http://www.ncbi.nlm.nih.gov/pubmed/21611761. Accessed 19 June 2013.

van Orsouw, N.J., Hogers, R.C.J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H. and van Eijk, M.J.T. (2007) Complexity reduction of polymorphic sequences (CRoPSTM): a novel approach for large- scale polymorphism discovery in complex genomes. PLOS One , 11 , e1172

Yamanaka N, Lemos N (2013) Resistance to Asian soybean rust in soybean lines with the pyramided three Rpp genes. Crop Breed …: 75–82. Available: http://www.scielo.br/scielo.php?pid=S1984-70332013000100009&script=sci_arttext. Accessed 9 October 2013.

Yamanaka N, Lemos N (2011) Soybean breeding materials useful for resistance to soybean rust in Brazil. Japan Agric … 45. Available: http://ainfo.cnptia.embrapa.br/digital/bitstream/item/51746/1/JARQ45.4.pdf. Accessed 7 August 2013.

Varshney , R.K. , Nayak , S.N. , May , G.D. and Jackson , S.A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding . Trends Biotechnol. 27 : 522 – 530 .

Wu X, Ren C, Joshi T, Vuong T, Xu D, et al. (2010) SNP discovery by high-throughput sequencing in soybean. BMC Genomics 11: 469. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3091665&tool=pmcentrez&rendertype=abstract.

Zhang G, Xu S, Mao W, Hu Q, Gong Y (2013) Determination of the genetic diversity of vegetable soybean [Glycine max (L.) Merr.] using EST-SSR markers. J Zhejiang Univ Sci B 14: 279–288.

2 Objetivos

O presente trabalho tem como objetivo identificar e validar marcadores SNPs na cultura da soja através da técnica de Genotipagem por sequenciamento, para serem utilizados no mapeamento de características simples e complexas. Os SNPs associados às características de interesse poderão ser utilizados para seleção de indivíduos desejáveis em programas de melhoramento através de seleção genômica ou seleção assistida por marcadores.

Objetivos Específicos:

- Sequenciar e genotipar cultivares e acessos selvagens de soja através da técnica de Genotipagem por Sequenciamento;
- Identificar e validar SNPs nas populações de soja;
- Validar a técnica de GBS para mapeamento de características de herança simples, como cor de flor, cor de pubescência e tolerância ao herbicida glifosato;
- Mapear por associação características complexas como resistência ao nematoide de galhas, M. incógnita.

3 Mapping single traits in soybean by Genotyping by Sequencing approach.

André L. de L. Passianotto [1,2]; Humira Sonah [3]; Marcelo F. Oliveira [2]; François Belzile [3]; Ricardo V. Abdelnoor [2]

[1] Graduate Program in Genetics and Molecular Biology, Londrina State University, Londrina, PR, Brazil
[2] Brazilian Agricultural Research Corporation, National Soybean Research Center (Embrapa – Soja), P.O. Box 231, Londrina, PR, Brazil
[3] Department of Plant Sciences and Institute of Integrative Biology and Systems (IBIS), Université Laval, Quebec City, Quebec, Canada G1V 0A6

Corresponding author: ricardo.abdelnoor@embrapa.br

3.1 Abstract

Soybean, with a production of more than 80 million tons per year, is one of the main commodities in Brazil, being very important for a positive trade balance exports. With a moderately complex genome, soybean adds a degree of difficulty in conducting genomic projects. However, with the advent of new tools for next-generation sequencing, the knowledge gained from the study of the genome of that oilseed has increased considerably in recent years. Despite the advances in sequencing technologies and the consequent cost reduction, the evaluation of a large number of individuals is still a challenge. Recently, Genotyping by Sequencing (GBS) approach has been used for a cost effective alternative for single nucleotide polymorphism (SNP) identification and its use on genetic mapping by association studies. The present study aimed to evaluate a set of soybean cultivars by GBS and validate this technique for mapping single traits in soybean. On this study, 12,303 SNPs were identified on 165 cultivars. These SNPs were used to map three single traits: glyphosate resistance (RR), pubescence color and flower color, that were mapped on chromosomes 2, 6 and 13 respectively. These results confirmed previous studies where those traits had been already mapped based on SSR markers and bi-parental populations. In total, 13 SNPs were associated with these traits, with qFDR > 0.001, showing the effectiveness of this technique on mapping specific traits in soybean.

3.2 Introduction

The cultivated soybean (*Glycine max* (L.) Merrill) was originally domesticated in China [1-2] and nowadays is widely grown across all continents [3]. Currently, the United States and Brazil are the main producers, accounting for 62% of world production in 2012/2013 [4]. Since the inception of intense commercial production, in the 20th century, soybean has been the object of intense selection, allowing its adaptation to a wide variety of environments.

During the last decades, the use of molecular markers, revealing polymorphism at the DNA level, has opened the way to numerous applications in plant breeding and genetics. Simple and complex traits have been mapped in plant genomes. Using both conventional breeding and biotechnology, the introduction of traits of commercial interest for soybean breeding has become possible with a high accuracy. Such tools have contributed significantly to an increased accuracy and precision of selection [5]. A very broad range of molecular markers, such as *restriction fragment length polymorphisms* (RFLP) [6], *amplified fragment length polymorphisms* (AFLP) [7], *random amplified polymorphic DNA* (RAPD) [8], *simple sequence repeats* (SSR) [9] and *single nucleotide polymorphisms* (SNP) [10] have been successfully used in plants.

Until recently, SSR markers had been the marker of choice in many plant species, including soybean. SSR markers are very frequent, randomly distributed throughout the genome, codominant, and have been widely used in studies of genetic diversity, germplasm characterization, construction of genetic maps and mapping of genes and quantitative trait loci (QTLs). The complete sequencing of the soybean genome [11] revealed that it is a moderately complex genome (~1.1 Gb) and more than 200,000

SSR loci has been identified so far [12]. Nonetheless, genotyping such SSR loci is not possible in a highly parallel fashion, as typically SSRs can only be analyzed a few at a time. For some applications, such as genome-wide association analysis, this represents an important limitation and alternative marker systems have been developed.

Next generation sequencing (NGS) technologies have allowed filling this gap. Large-scale sequencing efforts first opened the way for the initial discovery of SNPs in soybean, i.e. the description of variable base positions in the soybean genome. This information was then used to develop highly parallel SNP genotyping arrays such as the Universal Soybean Linkage Panel (1,536 SNP markers [13]) and, more recently, the SoySNP50K array (>47,000 SNPs [14]). An alternative strategy has been to use the power of NGS to simultaneously perform SNP discovery and genotyping. This is best exemplified by an approach termed genotyping by sequencing (GBS), where complexity reduction methods allow the massive sequencing of the same subset of the genome in many samples [15]. While GBS was first developed in maize and barley, an optimized protocol was recently reported for soybean [10]. The high degree of efficiency of this methodology allows studies on genomic scales, where a large number of high quality SNPs need to be generated at a low cost per sample.

This revolution in genotyping has now made it possible to densely cover the genome with tens of thousands of markers. In turn, this has made it possible to perform genome-wide association studies (GWAS) in soybean [16-17]. In this new approach [18], there is no need to establish a specific mapping population via labor-intensive crossing and advancing generations. Instead, readily available materials such as cultivars and breeding lines can be used to map the loci controlling traits of interest.

This study aimed to explore the feasibility of conducting a GWAS in a collection of Brazilian soybean lines using a large set of SNP markers obtained by GBS approach.

In a first step, the present work aimed at mapping three traits known to be controlled by one or very few genes: pubescence color, flower color and glyphosate tolerance.

3.3 Material and Methods

3.3.1 The plant material

Leaf tissue was collected from 165 soybean cultivars grown in a greenhouse. The seed were obtained from the Embrapa Soja germplasm bank and from commercial sources (Table S1). Six young plants were arranged in pots of 3 kg and younger trifoliate leaves of each plant were collected, frozen in liquid nitrogen and subsequently stored at -80°C.

The phenotypic data for flower color, pubescence color and the presence of the RoundUp Ready (RR) trait was obtained from the database of the National Service for Cultivar Protection (SNPC) and internal database information (M.F.O. personal information) (table 1).

Table 1: Summary of phenotypic characterization of soybean cultivars.

Trait	Phenotype	Individuals
Glyphosate tolerance	RR	51
	no_RR	114
Flower color	Purple	62
	White	103
Pubescence color	Gray	76
	Light brown	10
	Medium brown	40
	Brown	39

3.3.2 DNA extraction and library construction

Leaf samples (100mg) were ground in a mortar and DNA was extracted using the DNeasy Plant Mini Kit (Qiagen cat. No. 69106) as per the manufacturer's instructions. DNA was quantified using a Nanodrop 8000 spectrophotometer (Thermo Scientific; Wilmington, DE) and normalized to 10 ng/µl. The GBS libraries were prepared at the Plate-forme d'analyses génomiques (Université Laval, Quebec City, QC, Canada) essentially as per Elshire et al (2011), with the minor modifications described in Sonah

et al. (2013). Basically, DNA from the different genotypes was digested with *Ape*KI, ligated to adapters carrying a unique barcode, multiplexed to the libraries and sequenced on an Illumina GAIIx apparatus (McGill University-Genome Quebec Innovation Centre, Montreal, QC, Canada).

3.3.3 Data Analysis and SNP identification

Reads (108 bp long) were analyzed using a custom-designed pipeline implemented in perl language (IGST-GBS pipeline; J. Laroche, Université Laval, data unpublished). The processing of the Illumina sequence read data generated a “raw” catalog of SNPs that underwent further filtering. All heterozygous genotypes were removed and replaced with missing data; SNP loci and individuals with >20% missing data were removed. Finally, any missing data found at this stage were imputed using fastPhase [19] and only loci with a minor allele frequency (MAF) >0.05 were kept for further analysis.

3.3.4 Association mapping

Genome Wide Association Studies (GWAS) were conducted using a compressed mixed linear model [20] implemented in the Genomic Association and Prediction Integrated Tool – GAPIT [21]. Population structure was first determined by performing a principal component analysis (PCA) and the first six principal components were used in the ensuing analysis. Similarly, the genetic relatedness between the lines was captured in a similarity matrix. Marker-trait associations were then estimated using a mixed linear model (MLM) incorporating both population structure (6 PCs) and the similarity matrix. Associations were declared significant using a false discovery rate of 0.001.

3.4 Results

3.4.1 SNP discovery and SNP distribution

The sequenced GBS libraries yielded a total of 320 million 108-bp reads for the 165 cultivars. From these reads, a total of almost 55,000 “raw” SNPs were obtained and 12,303 SNPs remained after stringent filtering (described above). SNP markers were relatively evenly distributed on all 20 chromosomes, as shown in Figure 1. The number of SNP markers ranged between 1,084 (Gm18) and 373 (Gm01), for an average of 615 per chromosome.

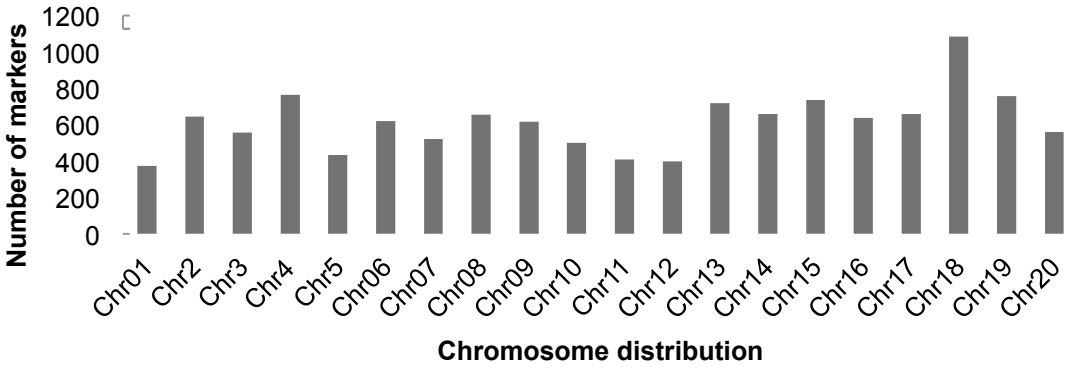

Figure 1. Distribution of the SNPs according to the soybean genome chromosome. The number of markers is correlated to its position according to the chromosome which occupies.

3.4.2 Population structure and genetic relatedness

The principal components are based on the collection of 12,303 SNPs. In this set of soybean lines, the six principal components explained 100% of the population structure, and when the first two principal components (PC1 and PC2) are plotted (Figure 2), a uniform scattering of the genotypes is observed without any clear subgroups being defined. The Brazilian material has a clear diversity considering the totality of the materials, opposing groups are evident, although there is the concatenation of some materials.

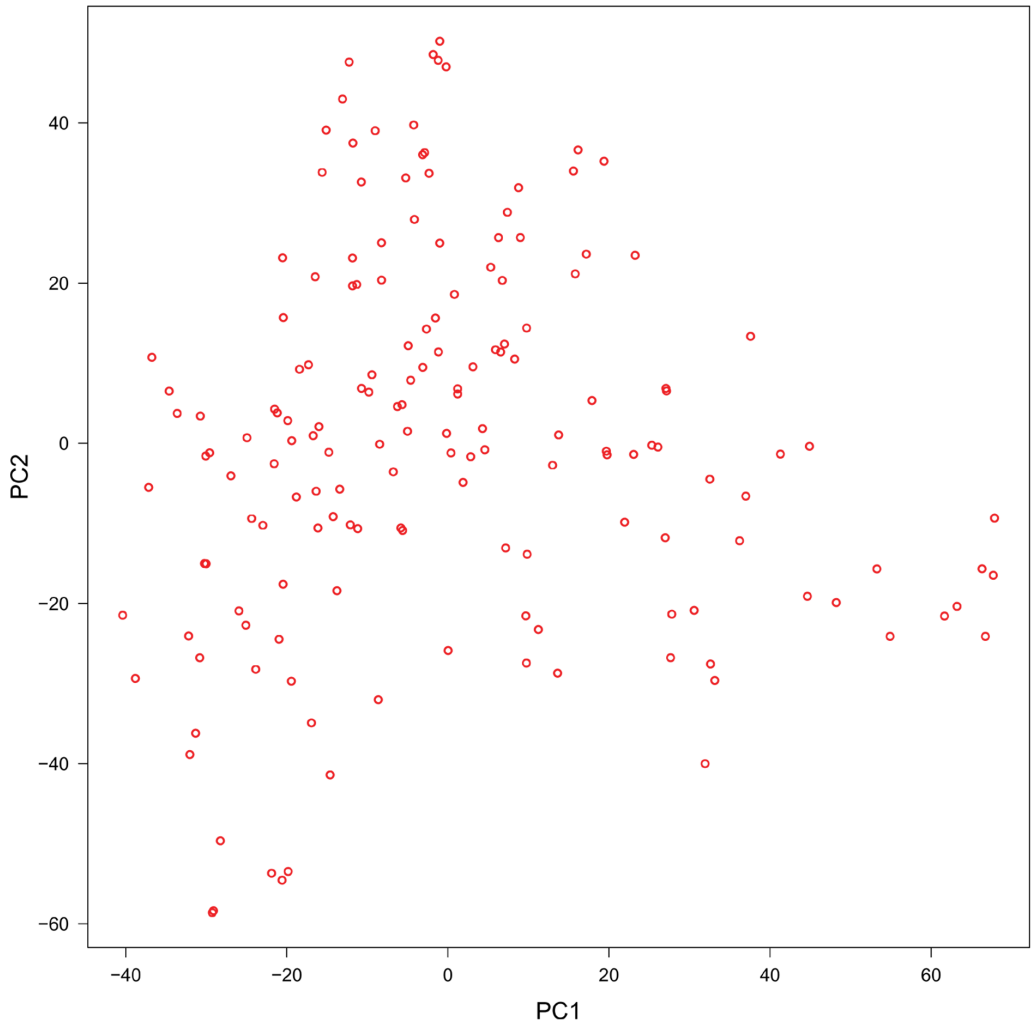

Figure 2. Genetic structure of the population according the Principal Component Analysis 1 and 2. Each genotype can be visualized by red circles. The position in the graphic refers to the genotype demonstrated by the individual within the collection of 12,303 SNPs.

3.4.3 Association Mapping Studies

In order to determine if the SNP data obtained via GBS provided sufficiently dense coverage of the soybean genome, we attempted to map the loci underlying three simply inherited traits: flower color, pubescence color and glyphosate resistance.

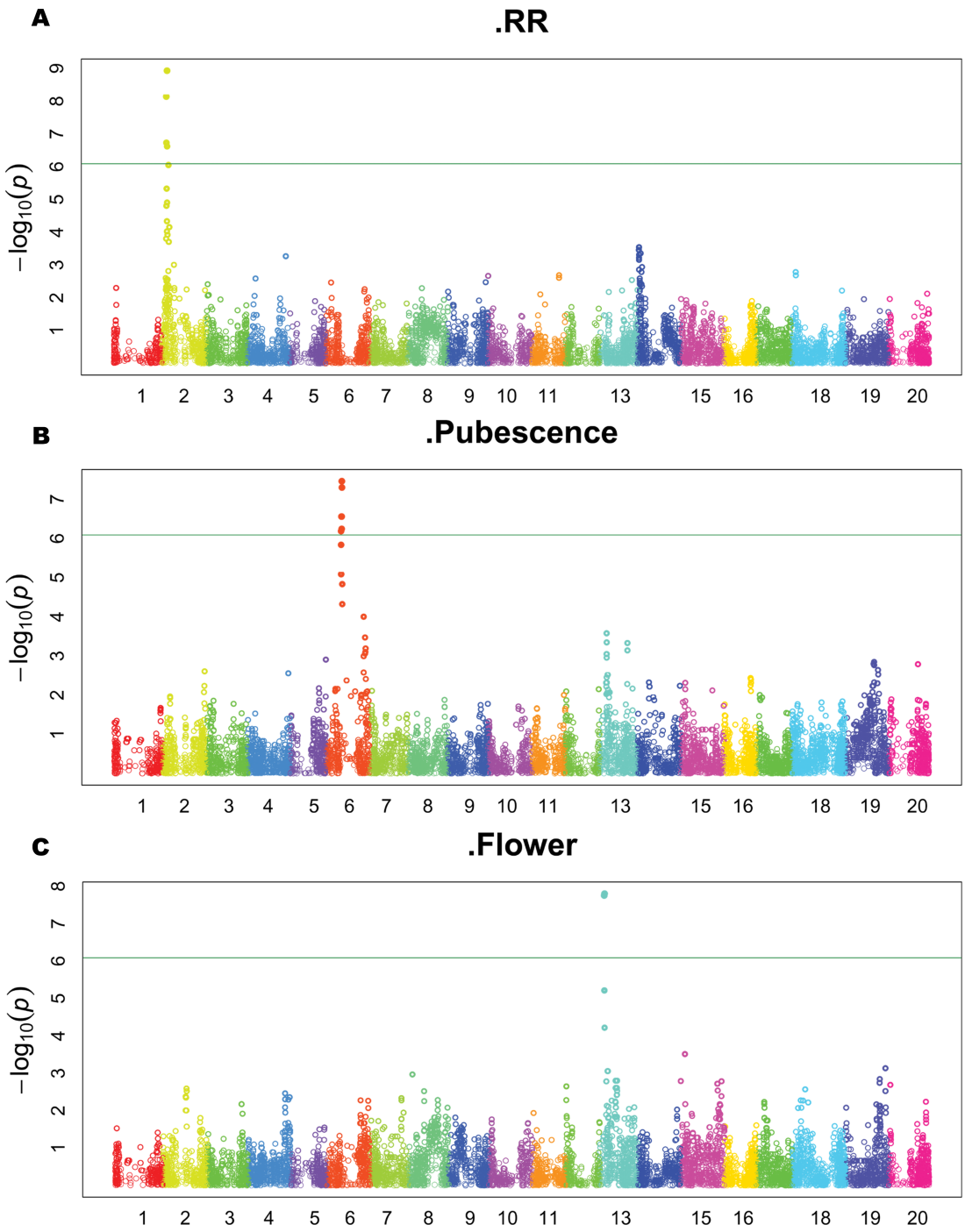

Figure 3. Manhattan plot displaying GWAS results for the three single soybean traits. A – Glyphosate Resistance (RR); B – Pubescence color; C – Flower color. Significant association was found for the SNPs with p-value above the threshold line.

Association mapping was performed for the three different traits using the filtered set of 12,303 SNPs and by accounting for population structure and genetic relatedness of the lines. As shown in figure 3, Manhattan plots were obtained for each of these traits

and significant marker-trait associations are listed in table 2. Among these associations, no marker-trait association exhibited a *p*-value inferior to 10^{-7}. Three different chromosomes were identified for each trait.

Table 2. List of the SNPs associated to the single traits. The SNPs are identified according to their location on the chromosome, their respective *p*-value, MAF and false discovery rate adjusting the *p*-value (FDR_Adj_P).

SNP Position	Chrom.	Trait	*p*-value	MAF	FDR Adj P
5887689	2	RR	1.18E-09	0.49	1.45E-05
5022121	2	RR	7.23E-09	0.48	4.45E-05
5131554	2	RR	1.86E-07	0.41	7.32E-04
5929459	2	RR	2.38E-07	0.41	7.32E-04
18515865	6	Pubescence	3.45E-08	0.46	3.06E-04
18637115	6	Pubescence	4.99E-08	0.49	3.06E-04
18404268	6	Pubescence	2.76E-07	0.49	5.65E-04
18405037	6	Pubescence	2.76E-07	0.49	5.65E-04
18066464	6	Pubescence	2.76E-07	0.49	5.65E-04
5019815	13	Flower	1.57E-08	0.45	7.21E-05
4695919	13	Flower	1.96E-08	0.30	7.27E-05
4565454	13	Flower	2.76E-07	0.30	9.21E-05
4800231	13	Flower	6.10E-07	0.38	1.87E-4

Glyphosate tolerance (RR) was mapped on chromosome 2 with the most significant association occurring with a SNP marker at position 5,887,689 (*p*-value = 1.18×10^{-9}) (figure 3a). Three other SNP markers located on one or the other side of this peak SNP also exhibited a significant association with this trait. Together, these four significantly associated SNPs encompassed a region of 900 kb. The locus controlling pubescence color was mapped on chromosome 6 (figure 3b) and five SNPs were significantly associated with the trait; these spanned an interval of 570 kb. The peak SNP in position 18,515,865 (*p*-value = 3.45×10^{-8}) is flanked by a set of three SNPs denoting a region of 232 kb in length. Flower color was found to be controlled by a locus situated on chromosome 13 (figure 3c). All the associated SNPs were located in an interval of 454 kb, with the peak SNP in position 5,019,815 (*p*-value = 1.57×10^{-8}).

From the 12,303 SNPs evaluated, 13 SNPs identified using via association analysis were found to be associated with the three traits (Table 2).

3.5 Discussion

The association mapping conducted in this study used a collection of SNPs discovered and validated in 165 soybean cultivars. The GBS approach allowed the identification of a total of 12,303 SNPs distributed across all 20 chromosomes, with an average of 615 SNPs per chromosome. In a previous study, 10,120 SNPs were identified in a set of eight soybean cultivars [10], showing the potential of this approach in identifying high quality SNPs.

Until the advent of GWAS, most of the studies aimed at mapping genetic loci, be they for simple or quantitative traits, relied on the progeny derived from a biparental cross and only yielded information that was valid for this study population. With GWAS, the genetic loci underlying traits can now be determined for a broad set of lines/cultivars, from the entire germplasm used in one breeding program, to all the germplasm in use in a country or even the world. The main obstacle to conducting GWAS in crop species has been the lack of sufficient marker coverage. With the advent of high-throughput genotyping platforms, this obstacle is being overcome.

One of the technologies that is used for SNP genotyping in soybean is the Illumina GoldenGate Genotyping Assay, where 384 to 3,072, SNPs can be assayed in different populations [22]. However, the SNPs need to be validated in a population of interest and can be reduced to a few hundred markers if the population is not highly polymorphic. The studies of genome-wide need the full amplitude of characteristic, the complexity reduction in conjunction with NGS provide the ability for GBS of increase the

population constantly making this technique beneficial for the breeding programs, with that is possible access the genome wide range contained in their studies.

Studies aimed at mapping traits controlled by a single genetic locus can, most of the time, be identified using SSR markers in a mapping population segregating for this trait [23]. The studies that identified the simple traits used in this paper relied on only a few dozen markers [23,24,25]. In contrast, in this work performed on a large number of unrelated soybean lines representing the Brazilian germplasm, a total of 12,303 SNP markers were used to precisely map the chromosomal regions controlling three simple traits. In all cases, the marker-trait associations were declared significant at a *q*FDR <0.001. Although there is some variation in the cutoff criteria used in the literature, other workers having used a less stringent value (e.g. *q*FDR = 0.1; [20-27]), the very stringent threshold used here suggests that only very tightly associated markers are being reported. Had we used a less strict value (*q*FDR = 0.1), a total of more than 20 SNPs would have been declared to show significant association with each trait. In the present study, as the location of the genes underlying each trait is already known, we can determine if the associated SNP markers identified in this work are indeed in close proximity to the causal gene.

The transgenic event 40-3-2 or RR trait [28] is important for agriculture in several countries because it reduces the production costs, mainly in situations where the pressure exerted by weeds is high. The use of this technology in the field has resulted in an increased efficiency of weed control and has reduced the amount of herbicides used [29]. The 40-3-2 event had been previously mapped on linkage group D1b (chromosome 2), using 53 SNPs, however many of these SNPs were false positives [25]. In our study, four SNPs were identified showing significant association to the RR trait, located on the same region previously described for the transgenic event 40-3-2.

All the SNPs showed high degrees of association with the trait. By selecting these SNPs, more than 88% of the genotypes were distinguished in our population. The markers associated to RR trait have a *p–value* ≤1.18 x 10^{-9} and this is similar to the one reported previously ($p < 1 \times 10^{-9}$) [25].

The studies conducted by Yang et al., in 2010, reported that the pigmentation of pubescence in soybean is controlled by a single dominant gene *T*. A dominant allele (*T*) confers brown color and its recessive allele (*t)* a gray color [26]. The difference between these two alleles is only a deletion of one nucleotide. The *T* gene shows the peculiarity of interacting with other genes (*I*, *R*, *O* and *W1*) that control the color of both the hilum and seed coat [26]. The *T* gene was sequenced and is given the identifier *Glyma06g*21920 (*T* or *F3'H* gene) as it lies on chromosome Gm06 between positions 18,534,682 and 18,541,273 [24,30]. Among the SNPs associated with pubescence color, none were located inside the coding region, but the two most significant SNPs (SNP 18,515,865 and 18,637,115) are located upstream and downstream of the gene, respectively, in an interval of approximately 115 Kb (Figure 4). This shows that the SNPs identified in this work correctly identified the chromosomal region controlling pubescence color, but also were located in very close proximity to the causal gene.

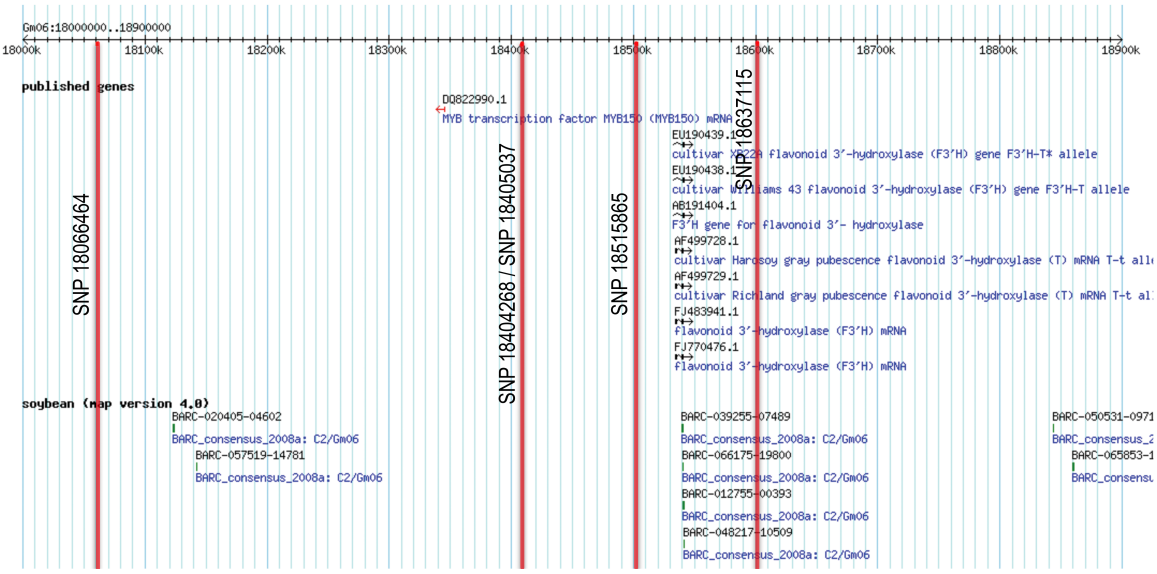

Figure 4. Genomic region containing the F3'H or T gene, responsible for soybean pubescence color. The figure was extracted from Soybase and the red line demonstrates the associated SNP.

Six genetic loci (*W1*, *W2*, *W3*, *W4*, *Wm*, and *Wp*) have been reported to control flower pigmentation in soybean [30]. The hydroxylation pattern of flavonoids plays an important role in the coloration of seed coats, flowers and pubescence of soybeans [31]. However, most of the cultivated soybeans have purple or white flowers, respectively conferred by the *W1* and *w1* alleles [26]. The flavonoid 3'5' hydroxylase (*F3'5'H)* was identified in previous studies as being associated with the *W1* locus on the basis of the analysis of a *F3'5'H* mutant [31]. The location of the W1 gene in soybean has been confirmed on chromosome 13, with SSR markers Satt348 and Satt160 [26] in accordance with the soybean genetic map [32]. According to Soybase, the *F3'5'H* gene is located on chromosome Gm13 between positions 4,552,711 and 4,557,278. In our study, the SNP most highly associated with flower color (SNP 5,019,815) is located 462 kb downstream of the *W1* gene. Interestingly, the SNP at position 4,565,454, located only 4.7kb downstream from the gene (Figure 5) and two other SNPs within 10 kb of the gene, were also identified as associated with this trait, but to a lesser degree. This

suggests that although these SNP markers are in closer physical proximity with the causal gene, they do not share as much linkage disequilibrium, possibly due to a different (more recent) mutational history.

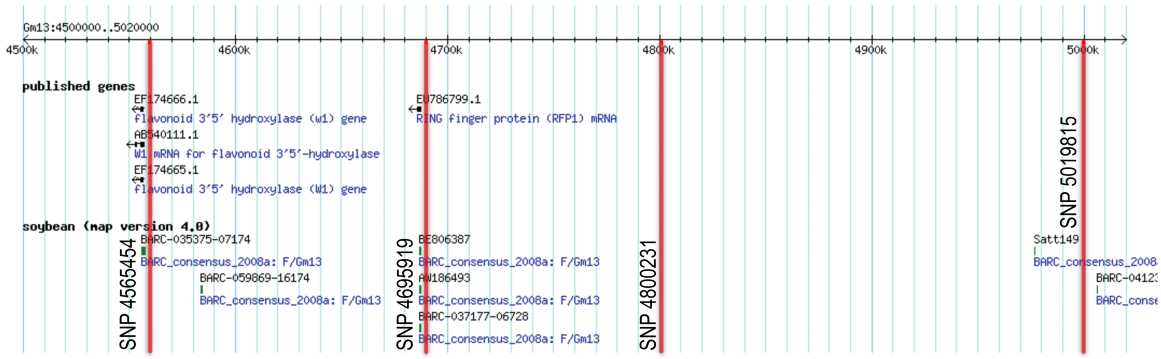

Figure 5. Genomic region of flavonoid 3'5' hydroxylase or W1 gene, responsible for flower color in soybean. The figure was extracted from Soybase and the red line demonstrates the associated SNP.

The highest SNPs associated with each of the traits explained an average of 80.3% of the observed phenotypic variation. Although one might have expected this proportion to be higher for such simple traits, some inaccuracies are possible and may occur because of differences between the material certified and the plant tissue used. Also, due to the physical distance between the markers and the target genes, we do not expect such associations to be perfect and explain all of the phenotypic variance.

With all the information obtained, we can conclude that few SNPs from the collection were strongly associated with the three single traits, on regions nearby, flanking the genes of interest. However, none of the SNPs was found inside the target genes. As the GBS approach contemplates the discovery of SNPs near to enzyme cleaved regions, it is possibly that the target genes does not contain *Ape*K1 sites. Finally, we can conclude that the GBS approach can be used efficiently to map specific traits in soybean and can be of great value for mapping new important agronomic traits.

3.6 Acknowledgments

The authors thank the Canadian Government for granting the scholarship entitled Emerging Leaders in the Americas Program (ELAP), Science Without Borders - CNPq by granting the PhD sandwich scholarship and CAPES by granting a PhD scholarship to ALLP. Approved for publication by the Editorial Board of Embrapa Soja as manuscript 04/2014

3.7 References

1. Fukuda Y (1933) Cytogenetical studies on the wild and cultivated Manchurian soybeans (Glycine L). Japanese Journal of Botany 6:489–506.
2. Vavilov NI (1951) Economic Botany: The origin variation immunity and breeding of cultivated plants. NewYork, Chron Bot Ronald Press. 364 p.
3. Chang RZ, Sun JY, Qiu LJ. The development of soybean germplasm in China. Crops 3:7–9,1998.
4. Companhia Nacional de Abastecimento – CONAB. Acomp. safra bras. grãos, Quarto Levantamento, Brasília, 1:67. Disponível em: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_01_10_15_07_19_boletim_graos_janeiro_2014.pdf.
5. Zhang G, Xu S, Mao W, Hu Q, Gong Y (2013) Determination of the genetic diversity of vegetable soybean [Glycine max (L.) Merr.] using EST-SSR markers. J Zhejiang Univ Sci B 14: 279–288.
6. Hisano H, Sato S, Isobe S, Sasamoto S, Wada T, et al. (2007) Characterization of the soybean genome using EST-derived microsatellite markers. DNA Res 14: 271–281.
7. Maughan PJ, Maroof MAS, Buss GR. (1996) Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor Appl Genet 93:574–579.
8. Abdelnoor, RV; Barros, EG; Moreira, MA. (1995) Determination of genetic diversity within Brazilian soybean germplasm using random amplified polymorphic DNA techniques and comparative analysis with pedigree data. Revista Brasileira de Genética, 18:265-273.
9. Akkaya, M.S.; Shoemaker, R.C.; Specht, J.E.; Bhagwat, A.A.; Cregan, P.B. (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. **Crop Science**, 35:1439-1445,
10. Sonah H, Bastien M, Iquira E, Tardivel A, et al. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8: 1–9.
11. Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, et al. (2010) Genome sequence of the palaeopolyploid soy- bean. Nature 463:178-183.

12. Song Q, Jia G, Zhu Y, Grant D, Nelson RT, Hwang EY, Hyten DL, Cregan P. (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. Crop Sci, 50:1950–1960.
13. Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matu- kumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor and Appl Gen 116:945–952.
14. Song Q, Hyten DL, Jia G, Quigley C V, Fickus EW, et al. (2013) Development and Evaluation of SoySNP50K , a High- Density Genotyping Array for Soybean. 8: 1–12.
15. Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):E19379.
16. Xu X, Zeng L, Tao Y, Vuong T, Wan J, et al. (2013) Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. Proc Natl Acad Sci USA 110: 13469–13474.
17. Bastien M, Sonah H, Belzile F. (2013) Genome wide association mapping of Sclerotinia sclerotiorum resistance in soybean with a genotyping by sequencing approach. doi: 10.3835/plantgenome2013.10.0030; Posted online 20 Dec. 2013
18. Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9: 166–177.
19. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
20. Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–U118.
21. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, et al. (2012) GAPIT: genome association and prediction integrated tool. Bioinforma 28 : 2397–2399.
22. Illumina - (http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn) accessed in December 2013.
23. Cregan, P.B.,T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. (1999) An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490.
24. Toda K, Yang D, Yamanaka N, Watanabe S, Harada K, Takahashi R. (2002) A single-base deletion in soybean flavonoid 3-hydroxylase gene is associated with gray pubescence color. Plant Mol Biol. 50:187–196.
25. Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK (2007) Molecular Markers in a Commercial Breeding Program. Crop Sci 47: S–154.
26. Yang K, Jeong N, Moon J-K, Lee Y-H, Lee S-H, et al. (2010) Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. J Hered 101: 757–768.
27. Hao D, Cheng H, Yin Z, Cui S, Zhang D, et al. (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. Theor Appl Genet 124: 447–458.
28. Padgette, S.R., K.H. Kolacz, X. Delannay, D.B. Re, B.J. LaVallee, C.N. Tinius, W.K. Rhodes, Y.I. Otero, G.F. Barry, D.A. Eichholtz, V.M. Peschke, D.L. Nida, N.B. Taylor, and G.M. Kishore. (1995) Development, identification, and characterization of a glyphosate-tolerant soybean line. Crop Sci. 35:1451–1461.
29. Menegatti, A.L.A.; Barros, A.L.M. (2007) Comparative analysis of production costs between conventional and transgenic soybean: a case study for the state of Mato Grosso do Sul Journal of Economics and Rural Sociology, v.45, n.1, p.163 -183.

30. Palmer, R.G., T.W. Pfeiffer, G.R. Buss, and T.C. Kilen. (2004) Qualitative genetics. p. 137–233. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, production, and uses. 3rd. ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI.
31. Zabala G, Vodkin LO. (2003) Cloning of the pleiotropic T locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3# hydroxylase. Genetics. 163:295–309.
32. Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB. (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet. 109:122–128.
33. Takahashi R, Stephen M, Hatayama GK, Dubouzet EG, Shimada N, Aoki T, Ayabe S, Iwashina T, Toda K, Matsumura H. (2007) A single-base deletion in soybean flavonol synthase gene is associated with magenta flower color. Plant Mol Biol. 63:125–135.

3.8 Support information

Supplementary information 01. **List of soybean cultivars used in this study.**

1	BRS216	56	BRSGO 8360RR	111	Embrapa1(IAS5RC)
2	BMX Potencia	57	BRSGO 8660	112	Embrapa4(BR4RC)
3	BMX Turbo RR	58	BRSGO Gisele	113	Embrapa48
4	BR 40 (Itiquira)	59	BRSGO LuzianiaRR	114	Embrapa59
5	BR1	60	BRSGO MINEIROS	115	EMGOPA 315 RR
6	BR16	61	BRSGO204[Goiania	116	EMGOPA313
7	BR36	62	BRSGO8061	117	EMGOPA316RR
8	BR37(Original)	63	BRSGOAmaralina	118	FMTMatrinxa
9	BRS 246RR	64	BRSGOChapadoes	119	FMTTucunare
10	BRS 270 RR	65	BRSGOEdeia	120	FT10(Princesa)
11	BRS 314	66	BRSGOIara	121	FTCristalina
12	BRS 316RR	67	BRSGOIndiara	122	Hardee
13	BRS 333 RR	68	BRSGOIpameri	123	IAC2
14	BRS 359	69	BRSGOJulianaRR	124	IAC5
15	BRS 360RR	70	BRSGOMineiros	125	IAC8
16	BRS 7860 RR	71	BRSGORaissa	126	Industrial
17	BRS 8160 RR	72	BRSGralha	127	M-soy 7211RR
18	BRS Baliza	73	BRSInvernada	128	M-SOY 8199
19	BRS Charrua	74	BRSMacota	129	M-SOY 8248
20	BRS Jiripoca	75	BRSMG 752 S	130	M-SOY 8787 RR
21	BRS SilvaniaRR	76	BRSMG 811 CRR	131	M-SOY5826
22	BRS TAURA RR	77	BRSMG 740S RR	132	M-SOY7501
23	BRS Tertulia	78	BRSMG 751 SRR	133	M-SOY8001
24	BRS TordilhaRR	79	BRSMG 850GRR	134	M-SOY8008
25	BRS184	80	BRSMG250[Nobreza]	135	M-SOY8336RR
26	BRS185	81	BRSMG760 RR	136	M-SOY8352
27	BRS213	82	BRSMG800A	137	M-SOY8411
28	BRS214	83	BRSMG810C	138	M-SOY8585
29	BRS217[Flora]	84	BRSMilena	139	M-SOY8866
30	BRS218[Nina]	85	BRSMS 750 SRR	140	M-SOY8870
31	BRS231RR	86	BRSNovaSavana	141	M-SOY8925
32	BRS232	87	BRSPetala	142	NK 3363
33	BRS233	88	BRSPirarara	143	NK 7059 RR
34	BRS239	89	BRSRaimunda	144	NK 7074 RR
35	BRS242 RR	90	BRSSambaiba	145	NK412113(VMAX)
36	BRS245	91	BRSSinuelo	146	OCEPAR10
37	BRS255RR	92	BRSTiana	147	OCEPAR3(Primavera)
38	BRS263[Diferente]	93	BRSTorena	148	OCEPAR4(Iguaçu)
39	BRS267	94	BRSValiosaRR	149	OCEPAR8
40	BRS268	95	CamposGerais	150	OCEPAR9
41	BRS282RR	96	CD 225RR	151	P98C81
42	BRS283RR	97	CD 226RR	152	P98Y51

43	BRS284	98	CD 243RR	153	Parana
44	BRS285RR	99	CD 247 RR	154	Pelicano
45	BRS295RR	100	CD201RR	155	Perola
46	BRS317	101	CD202	156	Planalto
47	BRS334RR	102	CD206	157	SantaRosa
48	BRS8460RR	103	CD208RR	158	TMG 103RR
49	BRSCaiaponia	104	CD214RR	159	TMG 121RR
50	BRSCandieiro	105	CD215RR	160	TMG108RR
51	BRSCarla	106	CD217	161	Tropical
52	BRSCeleste	107	CD219RR	162	UFVS2001
53	BRSCorisco	108	CD244RR	163	Uniao
54	BRSEstânciaRR	109	Cristalina	164	V.MAX RR
55	BRSGO 7960	110	Davis	165	Vicoja

4 Genome wide association study for resistance to the southern root-knot nematode (*Meloidogyne incognita*) in soybean

Passianotto, A. L. de L.[1,2]; Sonah, H.[3]; Bastien, M.[3]; Tardivel, A.[3]; Belzile, F.[3]; Ilquira, E.; Légaré, G[3]; Boyle, B. [3]; Jean, M. [3]; Binneck, E[2]; Dias, W.P. [2]; Marcelino-Guimaraes, F.C[2]; Oliveira, M.F.[2]; Abdelnoor, R.V.[2]

[1] Graduate Program in Genetics and Molecular Biology, Londrina State University, Londrina, PR, Brazil
[2] Brazilian Agricultural Research Corporation, National Soybean Research Center (Embrapa – Soja), P.O. Box 231, Londrina, PR, Brazil
[3] Department of Plant Sciences and Institute of Integrative Biology and Systems (IBIS), Université Laval, Quebec City, Quebec, Canada G1V 0A6

Corresponding author: ricardo.abdelnoor@embrapa.br

To be submitted in TAG - theoretical and applied genetics, 2014

4.1 Abstract

Soybean [*Glycine max* (L.) Merrill] is one of the most important traded commodities for the Brazilian economy. However, soybean cultivation is often affected by biotic and abiotic factors that prevent the crop from attaining its full yield potential. With the advent of new tools for next-generation sequencing, the genomic knowledge gained from the study of this major oilseed crop has increased considerably in recent years. In this study, we performed a genotypic characterization of 189 plant introductions (PIs) using genotyping-by-sequencing (GBS) approach and a phenotypic characterization for resistance/tolerance to the southern root-knot nematode, *M. incognita*, allowing to perform a genome-wide association study (GWAS) for this important trait. From 17,530 SNP markers identified and validated on this set of genotypes, only five were significantly associated with nematode resistance. Remarkably, all of these were located in a single, very small (12 kb) region on chromosome 10. This genomic region has previously been reported to contain possible candidate genes for nematode

resistance by QTL mapping in a biparental cross. Most of the lines (48 out of 58) with the highest level of resistance shared the haplotype composed of the alleles associated with resistance at these 5 SNP loci. Interestingly, 10 of the lines exhibiting a high level of resistance did not exhibit the “resistant haplotype” on Gm10, suggesting that these lines possess a different genetic basis for their resistance.

4.2 Introduction

Soybean [*Glycine max* (L.) Merrill] is widely grown in Brazil and holds a prominent place in the commodities market. In recent years, it has posted several successive records in terms of its contribution to the country’s trade surplus, and it is exported to a large number of countries (CONAB 2013). In order to reach crop yield potential in the field, the environmental conditions must be as close as possible to ideal. However, biotic and abiotic stresses frequently prevent the crop from reaching its full potential. Among these, the southern root-knot nematode (RKN), *Meloidogyne incognita* (Kofoid & White) Chitwood, is a worldwide problem often resulting in major yield reductions (Riekert and Henshaw 1998; Fourie et al. 1999; Sikora et al. 2005; Bridge and Starr, 2007; Fourie, et al., 2013; Xu et al, 2013). In Brazil, the species of *Meloidogyne* affecting soybean production are *M. incognita*, *M. javanica* and *M. arenaria* (Almeida et al., 2005). *M. incognita* predominates in large areas previously cultivated with coffee and cotton and these areas have a great potential for soybean crop production (Almeida et al., 2005). Outbreaks of infestations are described in areas where the occurrence of this pest is observed as underdeveloped soybean plants and yellowing leaves (Embrapa, 2011).

Resistance to *M. incognita* is usually described as a complex trait and biparental populations have traditionally been used for the discovery of QTLs through analysis with molecular markers (Li et al., 2001; Ha, et al., 2004; Fourie et al., 2008; Shearin, et al., 2009). In the pioneering work of Li et al. (2001), one major QTL on chromosome 10 and a minor QTL on chromosome 18, both derived from exotic germplasm (PI96354), have been reported to confer resistance. In the work of Fourie et al. (2008), the same region of chromosome 10 was reported as a minor QTL while a region on chromosome 7 was said to constitute a major QTL. Yet another QTL was identified by Shearin et al. (2009) on chromosome 6 near the T locus. All of these studies have provided a limited resolution due to the small number of recombination events and low marker coverage that are typical of such biparental QTL mapping studies (Li et al., 2001; Ha, et al., 2004; Fourie et al., 2008; Shearin, et al., 2009; Varshney, et al 2009). Recently, Xu et al. (2013) used low-coverage resequencing of 246 RILs derived from a Magellan x PI438489B to perform QTL mapping with over 100,000 SNP markers. This work resulted in the detection of a major QTL on chromosome 10, near the previously described region, and two other minor QTLs on chromosomes 8 and 13. This allowed the authors to identify two possible candidate genes on Gm10 (Glyma10g02150 and Glyma02160), both of which are thought to be involved in pectin metabolism.

Such QTL mapping in biparental populations is limited in two important aspects. Firstly, it can only examine the allelic diversity that is segregating between the two parents of the population and, secondly, the genetic resolution is limited as few recombination events are captured in such populations (Korte and

Farlow, 2013). In contrast, in a genome-wide association study, the allelic diversity present in a wider panel of unrelated accessions can be sampled and, because of the much longer time elapsed since the last common progenitor of these accessions, linkage disequilibrium is much less extended leading to a much improved precision (Jannink et al., 2010). In the case of accessions known to confer resistance to RKN, it would be important to determine if all such lines share a common genetic basis for resistance or whether different accessions differ in the QTLs that provide this resistance.

It is only recently, however, that the genotyping tools available to the soybean community have provided the high density of marker coverage required to make GWAS possible. Although low-coverage re-sequencing provides the most complete description of the polymorphisms that exist within a set of germplasm, such data are currently available for only a limited number of soybean genotypes (Lam et al., 2010; Li et al., 2013), and remain quite costly to perform on hundreds of lines. SNP genotyping platforms provide an alternative. The SoySNP50K array developed by Song et al. (2013) allows one to interrogate ~47,000 SNPs, a portion of which will be informative within a specific set of accessions. An alternative approach, termed genotyping by sequencing (GBS) has been used in order to simultaneously identify and score SNPs on large populations of plants (Elshire et al., 2011). Recently, an efficient GBS protocol for soybean has been described (Sonah et al., 2013) and has successfully been used to perform a GWAS of resistance to white mold (Bastien et al., 2014).

The objective of this study was to identify SNPs in a set of 194 soybean accessions, including many that are known sources of resistance to the RKN, with the use of a GBS approach and to perform a GWAS to uncover the genomic regions containing genes conferring resistance/tolerance to RKN.

4.3 Material and methods

4.3.1 Plant tissue

A set of 189 plant introductions (PIs) and five soybean cultivars (Supporting Information 01), was used in this study. Three genotypes (PI 595099, BRS 282, CD 201) were used as resistant checks and three others (Santa Rosa, BRS Celeste and Embrapa 20 (Doko RC)) served as susceptible checks. The seeds were obtained from the Active Germplasm Bank of Embrapa, located at the National Soybean Research Center in Londrina, PR, Brazil. Ten seeds of each genotype were grown in a greenhouse. After 20 days, trifoliate leaves from 6 young plants were collected in bulk, frozen in liquid nitrogen and stored in a -80°C freezer. The leaf samples were ground to a fine powder and stored until DNA isolation.

4.3.2 Nematode resistance assay

Six plants of each genotype were evaluated for nematode resistance. The plants were grown in a greenhouse in plastic tubes containing 500 cm^3 of a mixture of autoclaved soil and sand. The plants were kept under 16 hours of daylight and subsequently supplemented with 600W high-pressure sodium lamps (Light Systems PL). At the V2 stage, a suspension of 5,000 eggs of *M. incognita* was inoculated to each tube. Weekly, 80 ml of nutrient solution (Hoagland & Arnon, 1950) was used to fertilize the plants. Thirty days after inoculation, roots of each plant were scored on a scale of 1 to 5 based on the abundance of galls, where 1 means a highly resistant plant and 5 a severely diseased plant. Nematode reaction was assessed on 6 plants per genotype and a nematode

score was obtained by calculating an average for each line. Scores between 1.0 and 1.8 were deemed to define the resistant class and scores between 3.8 and 5.0 were taken to represent susceptible genotypes .

4.3.3 DNA extraction and GBS library preparation

The DNA of each sample was extracted with the DNeasy Plant Mini Kit Kit (Qiagen), and subsequently quantified using a Nanodrop 8000 spectrophotometer (Thermo Scientific, Wilmington, DE). The samples were then diluted to 10 ng/ul. GBS libraries were then constructed according to the protocol described by Elshire et al. (2011), as modified by Sonah et al. (2013) to include a 2-bp (AC) selective amplification step. Amplicons were pooled to form essentially two (100-plex) GBS libraries and each was sequenced on a single lane of an Illumina HiSeq2000 apparatus (McGill University-Genome Quebec Innovation Centre, Montreal, QC, Canada).

4.3.4 Pipeline for SNP identification

Illumina sequence read processing, mapping, SNP/indel calling and genotyping was performed using the IGST-GBS pipeline (Sonah et al., 2013). Any heterozygous calls were replaced with missing data and only SNPs with less than 20% missing data were kept. SnpEff (http://snpeff.sourceforge.net) was used to annotate SNPs in terms of their position within the genome (intergenic, intron or exon) and to predict the impact of variant positions on the amino acid sequence of predicted proteins (synonymous or non-synonymous). Indels were not used in the downstream analyses. Imputation of any residual missing data was performed using fastPHASE 1.3 (Scheet & Stephens, 2006). For association

mapping a minor allele frequency (MAF) of ≥ 0.05 was used. Graphical genotype visualization was performed using Flapjack (Milne et al., 2010).

4.3.5 Population structure and genetic relatedness

The population structure and kinship matrix was created using the TASSEL program (Bradbury, et al., 2007). A set of SNPs adjusted in MAF ≥ 0.05, (17,530 SNPs) served as the basis for population structure. On the basis of the Scree plot (supplementary figure 01), the first eight principal components were used to capture the population structure. The kinship matrix was generated using TASSEL to calculate a distance matrix.

4.3.6 Association mapping

Marker-trait associations were calculated with GAPIT (Lypka, et al., 2012). A general linear model supported solely by phenotypic and genotypic data did not account the population structure and genetic relatedness between lines (naïve model). The P matrix resulting from the first eight principal components (PC) was used in addition to a kinship matrix (K) and a compressed mixed linear model (CMLM). Marker-trait associations were declared significant using FDR-adjusted *p*-values with the threshold set at 0.001.

4.4 Results

4.4.1 Phenotypic evaluation

The 194 soybean genotypes were evaluated for nematode resistance in a greenhouse assay (Figure 1; Supporting information 01). The three resistant checks (PI595099, BRS 282 and CD 201) all received the lowest scores (1.0, 1.2 and 1.2, respectively) while the susceptible checks (Santa Rosa, BRS Celeste and Embrapa 20 (Doko RC)) all had the highest possible score (5.0). Among the other accessions, 58 genotypes received scores between 1.0 and 2.0 (indicating a high level of resistance) and the remaining 136 genotypes exhibited scores ranging between 2.2 and 5.0 (Figure 1). Thus, this population of soybean accessions provided a fairly equal number of lines in each of the broad nematode reaction classes shown in Figure 1.

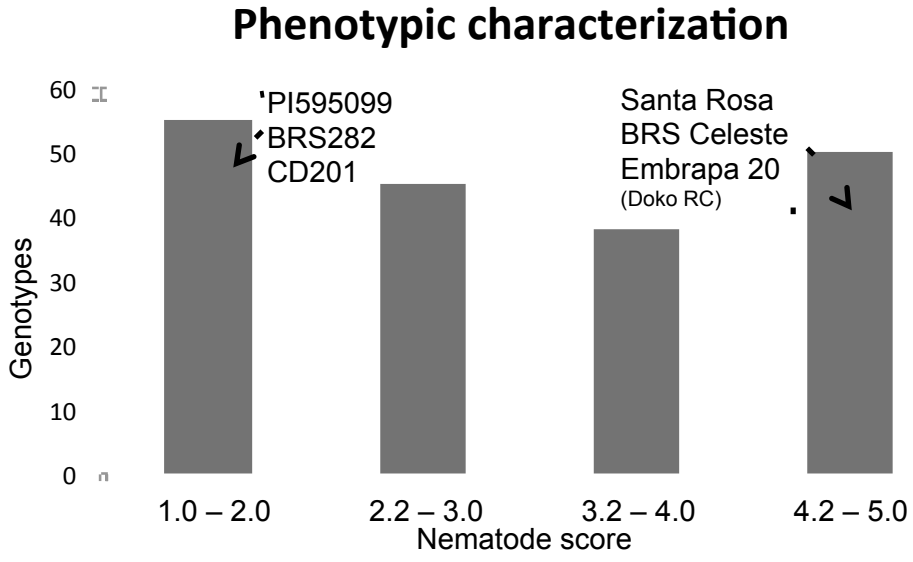

Figure 1. **Reaction of 194 soybean accessions to *M. incognita***. Distribution of mean nematode scores (1-5) obtained from six plants per accession.

4.4.2 Marker discovery, distribution and population structure

The two HiSeq lanes on which the GBS libraries were sequenced generated a total of 395 millions of 100-bp reads for an average of 2.03 million reads per sample. After running the IGST SNP-calling pipeline and eliminating SNPs with > 20% missing data, a total of 40,654 SNPs were identified on the 20 chromosomes, while a limited set of 291 polymorphisms mapped to the small contigs that remain unassigned to a chromosome. A total of 1,716 indels (4.2% of polymorphisms) were composed of 725 insertions and 991 deletions, but these polymorphisms were not used in the following analyses.

The SNP markers were distributed proportionately over the 20 soybean chromosomes with the largest chromosome (Gm18, 62 Mb) having the largest number of SNPs (3,046) and the shortest (Gm16, 37 Mb) exhibiting the smallest number of SNPs (1,869). As for the distribution of markers within coding vs non-coding regions, 18.1% of the SNPs resided in exons, 30.8% were in intergenic regions, 16.1% in upstream regions, 20.4% in downstream regions, 12.4% in introns and 2.6% in UTR regions. As for the type of mutations leading to these SNPs, 24,994 changes were transitions and 14,235 were transversions, leading to a ratio of transitions to transversions of 1.75.

For the purpose of performing a GWAS, the catalog of SNPs was further refined to keep only those markers that could be considered common (MAF ≥ 5%), resulting in a final list of 17,530 SNPs. As illustrated in Figure 2, the number of SNPs per chromosome ranged between a minimum of 468 (Gm12) and a maximum of 1,424 (Gm18).

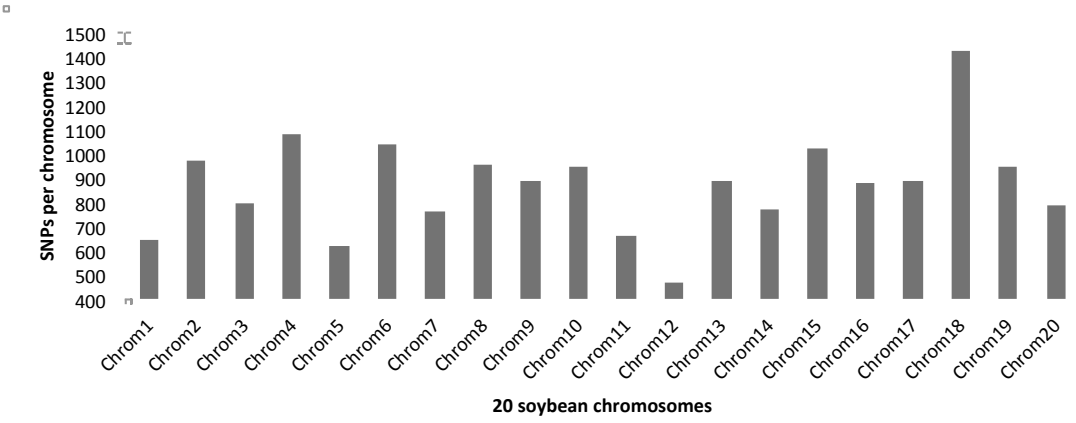

Figure 2. **Distribution of common SNPs on the 20 soybean chromosomes**. The graph shows the number of SNPs with a MAF ≥ 0.05 on each chromosome.

The first eight principal components (PCs) were used to produce a P matrix to account for population structure in the models used to estimate marker-trait associations. Collectively, these 8 PCs explained 33% of the variance, with PC1 alone explaining 10.7% of the variance. As shown in Figure 3, these accessions were widely dispersed with some clearly defined subgroups being apparent (circled). Within the range of markers used, the Brazilian cultivars lined up close to each other (blue circles). Filled circles identify resistant materials and are arranged around all genotypes set.

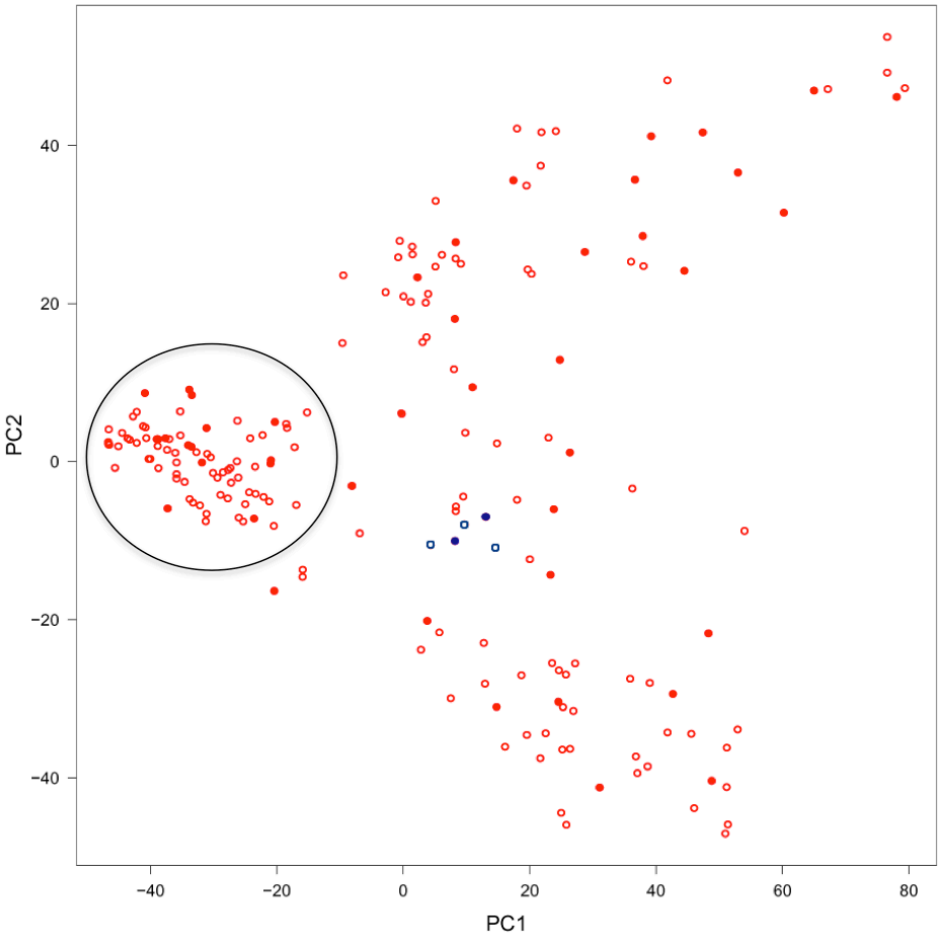

Figure 3. **A population structure sampled by PCA 1 and PCA 2 with the highest EigenValues.** The principal component, assist in the evaluation of the population structure of the population under test. Filled circles indicate resistant materials, red circles indicate PIs and blue circles cultivars.

4.4.3 Association mapping for nematode resistance

GAPIT was employed to perform the genome-wide association scan using a mixed linear model accounting for both population structure (P) and genetic relatedness (K) of the lines. The proportion of marker-trait associations with an observed p-value < 0.01 was 1% suggesting that the model has resulted in very

little, if any, inflation in the number of significant associations. Overall, this model performed very well as the cumulative distribution of observed p-values increased linearly (figure 4).

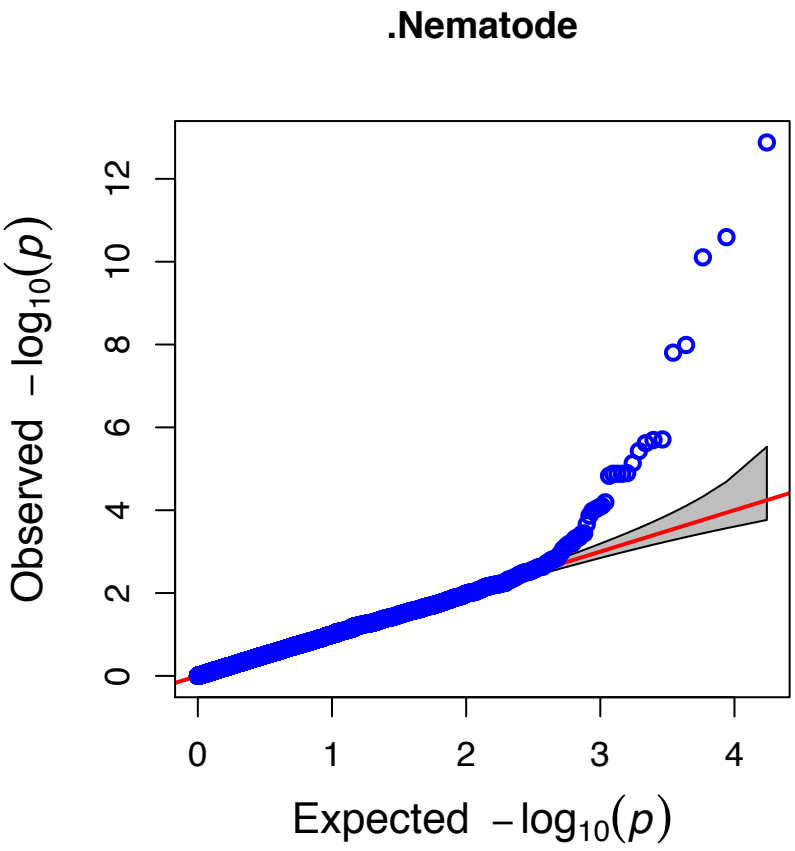

Figure 4. **Quantile-quantile plot (QQplot) of *P*-values.** The Y-axis is the observed negative logarithm of the *P*-values, and the X-axis is the expected negative logarithm of the *P*-values under the assumption that the *P*-values follow a uniform [0,1] distribution. The gray line show the 95% confidence interval for the QQ-plot under the null hypothesis of no association between the SNP and the trait.

As illustrated in Figure 5, the genomic scan resulted in a very low level of "basal" association (*p*-value > 10^{-3}) throughout the genome with the sole exception of a single region on Gm10, where a total of five markers in close proximity exhibited a very high degree of association.

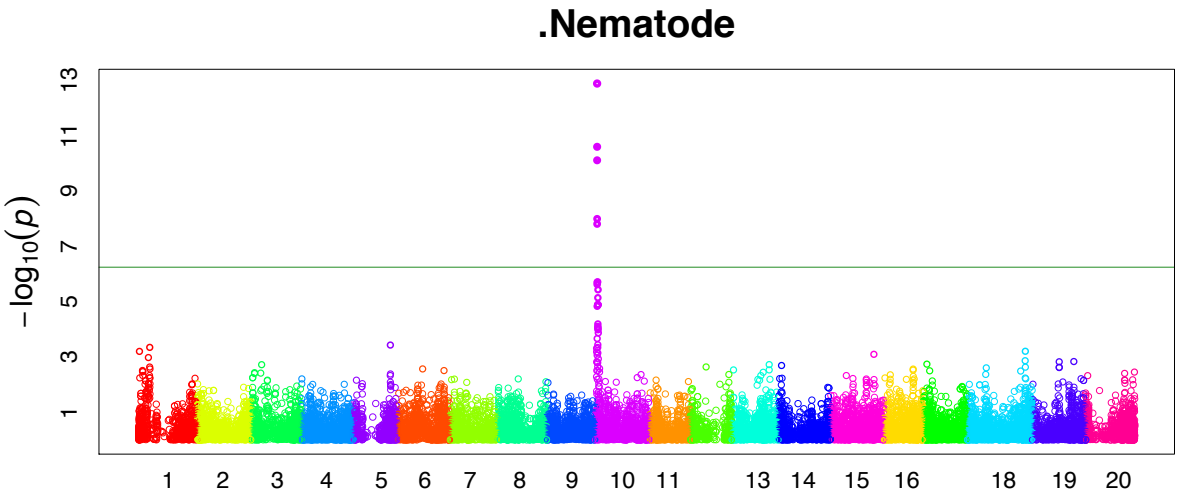

Figure 5. A **Manhattan plot illustrating genome wide association for resistance to *Meloidogyne incognita*.** Significant association was found for the SNPs with p-value above the green threshold line.

As shown in Table 1, the five significantly associated SNPs define a region spanning 12 kb, between SNP markers 1,502,515 and 1,514,717. The first of these showed the highest degree of association, which reached a *p*-value of 1.32 x 10^{-13}; this corresponds to a very low probability of being a false positive (2.3 x 10^{-9}). All SNPs within this region exhibited very highly significant *p*-values (< 1.57 x 10^{-8}), and had a very low probability of being a false positive (< 5.47 x 10^{-5}). The minor allele frequencies for these loci were relatively high, ranging between 0.22 and 0.35. Therefore, allelic effects could be estimated on fairly large numbers of individuals, ranging between 43 and 64 soybean accessions for each minor allele. The allelic effect estimated for the most significantly associated marker was 0.98, indicating that the allelic status at this locus had a very significant impact on the nematode score. The other significant markers had allelic effects ranging between -0.85 and 0.87. Accordingly, the proportion of phenotypic variance explained by these markers ranged between 25 and 40%.

Table 1. **The five SNPs associated with resistance against *Meloidogyne incognita* nematode.** All the markers are contained on chromosome 10, the position of peak marker on the physical map can be found in the table besides values of *p*-value, MAF - indicates whether minor allele provides increased resistance or susceptibility and the false discovery rate adjusted the *p*-value.

Position	Alleles	*p*-value	MAF	Allelic Effect	Rsq with SNP	FDR values
1502515	G/A	$1.3\ 10x^{-13}$	0.33	0.98	0.40	$2.3\ 10x^{-09}$
1505931	C/A	$2.5\ 10x^{-11}$	0.33	0.87	0.33	$2.2\ 10x^{-07}$
1505789	C/T	$7.8\ 10x^{-11}$	0.35	-0.85	0.32	$4.5\ 10x^{-07}$
1514717	A/T	$1.0\ 10x^{-08}$	0.22	0.83	0.26	$4.4\ 10x^{-05}$
1514707	G/A	$1.5\ 10x^{-08}$	0.22	-0.83	0.25	$5.4\ 10x^{-05}$

The three most significant SNPs (Table 1) were all located within a single predicted gene (*Glyma10g02140*), which putatively codes for a protein with two domains, one with similarity to a pectinesterase (PEC) and the other to a pectin methylesterase inhibitor (PMI). The peak SNP resides in the second intron and the two other associated SNPs are within the first exon. These two nucleotide changes are non-synonymous and result in two missense changes in the predicted amino acid sequence. As this predicted gene lacks a stop codon in the current genome annotation, lies in a region with many other similar genes and does not seem to be transcribed, it is possibly a pseudogene. The last two associated SNPs lie within in an exon of another gene, *Glyma10g02160*, which codes for a protein with the same two domains as *Glyma10g02140*. Again, the associated SNPs are non-synonymous and result in two changes in the predicted amino acid sequence, one silent and the other a missense. Here again, there is almost no evidence that this gene is transcribed.

The 194 soybean accessions could be classified based on the haplotype for the five most relevant SNPs associated to RKN resistance (Table 2). Most of the individuals fall in the resistant (ATATA) and the susceptible (GCCAG) haplotypes. Fifty-eight individuals were classified as resistant (scores between 1.0 and 2.0),

and 48 have the resistant “ATA” haplotype and ten have the haplotype “GCC” for *Glyma10g02140* (Supplem. Information 02). When considered the haplotype for the five SNPs, 41 out of 58 resistant individuals carry the resistant haplotype (ATATA). For the individuals classified as moderately resistant (scores between 2 and 3), and susceptible (scores between 3 and 5), most of them contain the susceptible haplotype (121 out of 136). Two indifividuals have the “ATAWA” haplotype, where the W is an indefined nucleotide (A or T). Based on the linkage between the SNPs, the two individuals may have T on the W position (Table 2).

Table 2. **The haplotype for five SNPs associated with the phenotypic evaluation for *Meloidogyne incognita* nematode.** All the SNPs with blue letters came from *Glyma10g02140*, and the red letters from *Glyma10g02160*.

	RKN score				
Haplotypes	1.0 - 2.0	2.0 - 3.0	3.0 - 4.0	4.0 - 5.0	Total
ATA TA	41	7	1	3	52
ATA AG	7	2	-	-	9
ATA WA	-	2	-	-	2
GCC AG	10	47	26	48	131
Total	58	58	27	51	194

4.5 Discussion

4.5.1 SNP discovery and distribution

Genotyping-by-sequencing has been considered an approach whose efficiency and cost effectiveness has been demonstrated in different studies (Elshire et al., 2011; Poland et al., 2012; Fu, 2012; Lu et al., 2012, Bastien, et al., 2013). A wide coverage on the genome and the multiplex possibility, make this approach a very attractive for SNP discovery in a set of large number of individuals. On this study, 40,654 high quality SNPs were identified over the 20

soybean chromosomes. On average, a SNP was detected every 23.3 kb. In a similar study, involving eight soybean lines, 10,120 SNPs were identified, resulting in a frequency of a SNP every 100 kb (Sonah et al., 2013). This difference may be mainly due the number of individuals being evaluated.

For association analysis purposes, 17,530 SNPs with MAF ≥ 5% were identified, validated and scored in the population, resulting an average of 873 SNPs per chromosome, ranging from 468 SNPs on chromosome 12 to 1424 on chromosome 18. Coincidently, the GM12 is one of the smallest (40Mb) and the Gm18 the largest (56Mb) ones, indicating a fairly similar frequency on the chromosomes, with a few exceptions. Although Gm12 has a similar size of Gm16 (37 Mb), the number of SNPs on Gm16 was approximately twice of Gm12. The location and number of SNPs per chromosome is strictly linked to the restriction sites. Nevertheless *Ape*K1 enzyme was shown to have a uniform distribution over the soybean genome, and regions surrounded by genes were favored instead regions containing repetitive information (Sonah, et al., 2013). Probably the Gm12 has lower informative region than Gm16.

4.5.2 Association Mapping for RKN resistance

Population structure was analyzed based on PCA using the SNPs that were genotyped in all 194 genotypes. As expected, the five soybean cultivars used as resistant and susceptible checks, were all located in a tight cluster. On the other hand, most of the PIs were spread over the plot, with no clear clustering. On the other hand, although most of the PIs show to be much more diverse than the elite Brazilian lines, a small defined cluster was formed suggesting that some of these accessions are more closely related. However, this group appears not to have

any link with the RKN resistance, since most of the lines that carries resistance genes are distributed throughout the plot.

With the sole exception of the recent work by Xu et al. (2013), all of the previous QTL studies on *M. incognita* resistance were performed with SSR markers, and most of these studies (Li et al., 2001, Ha, et al., 2004; Shearin, et al., 2009; Fourie et al., 2008) were conducted with relatively few markers. As such, each of these studies has been limited to exploring the QTL segregating in only one pair of lines. Although these have been useful and have repeatedly indicated that an important determinant of RKN resistance is found on Gm10, this is the first GWAS performed in soybean for this trait. Due to a set of 17,530 SNPs obtained via GBS, we were able to very precisely define a 12-kb interval on Gm10 that was very significantly associated with this trait. On this interval, three SNPs were found within *Glyma10g02140* and two on *Glyma10g02160*.

The CMLM approach (Zhang, et al., 2010) used in this work proved to be extremely successful at controlling for population structure and genetic relatedness. The PC +.K are show important in order to control false associations (Bastien, et al 2013) and the power of these two variables combined with the threshold supplied by GAPIT makes the result of this mapping very consistent. Based on Q-Q plot (Figure 4), constructed with a 17,530 markers set, only 167 SNP markers had a p-value < 0.009, and 99% of the markers were identified on the line or near for false positive SNP discovery.

The *Glyma10g02140* has been previously identified in a fine mapping study and was postulated as a candidate gene controlling the resistance to *M. incognita* (Pham et al., 2013). Xu et al. (2013), using a recombinant bin map approach, also found on this same region, five different genes as candidates for *M. incognita*

resistance. However, two of them were not considered due absence of start codon (*Glyma10g02140* and *Glyma10g02170*) or stop codon (*Glyma10g02140*).

One of the domains on these two gene models is a protein involved in the process of cell wall break down, where the pectinesterase catalyzes the esterification of pectin into pectate and methanol (Cosgrove 1997). Pectin is one of the main components of the plant cell wall and pectinesterase has been shown to play an important role in cell wall metabolism during fruit ripening (Hunter, et al., 2011). One the most frequently up regulated genes reported when occurs the cell wall penetration by nematode is the pectinesterase (Jammes, et al., 2005; Barcala, et al., 2010). Ibrahim, et al. (2011) showed that in *M. incognita* infected Arabidopsis and soybean plants, pectinesterase is one of the genes with the highest degree of significance.

4.5.3 Haplotype evaluation

The five SNPs associated with the RKN trait allowed the identification of four different haplotypes. Based on phenotypic evaluation, 58 individuals were considered resistant, with infection score lower than 2.0. These individuals presented three possible haplotypes “ATATA” (41), “ATAAG” (7) and “GCCAG” (10). If we consider only the haplotype of ATA (SNPs present on *Glyma10g02140*), 48 genotypes have the “ATATA” and “ATAAG” haplotype and carry the source of resistance on chromosome 10.

The remaining ten individuals, carry the susceptible haplotype (GCCAG), but were considered resistant. Among these individuals, the PI595099 was described as one important source of nematode resistance (Beneventi et al., 2013) and was used as one of the resistant checks. The other cultivars used as

resistant check, BRS 282 and CD 201, contains the haplotype “ATATA”, as most of the resistant individuals on this population. This leads us to believe that some resistance sources can carry resistance genes on other genomic region on soybean. Several other regions have been already identified as carrying QTLs for RKN resistance, as on the Gm06 (Shearin et al., 2009), Gm07 (Fourie et al., 2008), Gm08 (Xu et al., 2013), Gm13 (Xu et al., 2013) and Gm18 (Li et al., 2001). Taking into account the size of this subset of resistant individuals and the large number of QTLs conferring resistance described, it is possible that these other regions could not be detected with good association level. These additional resistance genes may also have not been detected because of the allelic frequency adjusted for 0.05. In this case, rares alleles were discarded in this association study. Then, the power of detection in the range of genomic features, might not be enough when the size of the population is not large enough to sample the trait.

4.6 Concluding remarks

One of the advantages of using GBS to perform a GWAS study is the possibility of analyze a broad population at low cost per sample. In this study, The association mapping analysis detected one region on chromosome 10 linked with high level of association and defined the shortest interval hitherto described for RKN. Although we were unable to detected a new QTL, the accuracy displayed by this technology is highly recommended for any other trait study in soybean.

The collection of SNP markers established on this study can supply the breeding programs with an informative database. This approach allows studies at

genomic scale, with high quality SNP data and a low cost per sample, making this technique an interesting alternative for mapping agronomic traits and genetic diversity on large populations, especially in the discovery of other sources of resistance.

4.7 Reference

Almeida, A.M.R.; Ferreira, L.P.; Yorinori, J.T.; Silva, J.F.V.; Henning, A.A.; Godoy, C.V.; Costamilan, L.M. & Meyer, M.C. 2005. Doenças da soja (Glycine max). In: Kimati, H.; Amorim, L.; Rezende, J.A.M.; Bergamim Filho, A. & Camargo, L.E.A. (Eds.) Manual de Fitopatologia. Vol. 2. Doenças das plantas cultivadas. 4a ed. Ceres. Piracicaba-SP. pp. 569-588

Barcala M, Garcı´a A, Cabrera J, Casson S, Lindsey K, Favery B, Garcı´a-Casado G, Solano R, Fenoll C, Escobar C (2010) Early transcriptomic events in microdissected Arabidopsis nematode- induced giant cells. Plant J 61:698–712

Bastien, M. Sonah, H. Belzile, F. (2013) Genome wide association mapping of Sclerotinia sclerotiorum resistance in soybean with a genotyping by sequencing approach. The Plant Genome: Posted 20 Dec. 2013; doi: 10.3835/plantgenome2013.10.0030

Beneventi MA, da Silva OB, de Sá MEL, Firmino AAP, de Amorim RMS, et al. (2013) Transcription profile of soybean-root-knot nematode interaction reveals a key role of phythormones in the resistance reaction. BMC Genomics 14: 322.

Bridge J, Starr JL. 2007. Plant nematodes of agricultural importance. Boston: Academic Press. p. 19–32.

Companhia Nacional de Abastecimento – CONAB. Acomp. safra bras. grãos, Quarto Levantamento, Brasília, 1:67. Disponível em: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_01_10_15_07_19_boletim_graos_janeiro_2014.pdf.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at rhg1 mediates nematode resistance in soybean. Science 338: 1206–1209.

Cosgrove DJ (1997) Assembly and enlargement of the primary cell wall in plants. Annu Rev Cell Dev Biol 13:171–201

Close T, Bhat P, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson J, Wanamaker S, Bozdag S, Roose M, Moscou M, Chao S, Varshney R, Szucs P, Sato K, Hayes P, Matthews D, Kleinhofs A, Muehlbauer G, DeYoung J, Marshall D, Madishetty K, Fenton R, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Howie Smith, O.S., Tingey, S., Morgante, M. and Rafalski, A.J. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3, 19 , 2002

Elshire, R. J.; Glaubitz, J. C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E. S.; Mitchell, S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6, e19379, 2011

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária. 2010. Cultivares de soja de Minas Gerais e região central do Brasil. Safra 2010/2011. Londrina, PR. Accessed in january 2013 :. http://www.cnpso.embrapa.br/download/cultivares/Soja_2010-11MG.pdf.

Edwards, D.; Forster, J. W.; Cogan, N. O. I.; Batley, J. ; Chagné, D.. (2007) Single Nucleotide Polymorphism Discovery. In Oraguzie, NC; Rikkerink, EHA; Gardiner, SE; Silva, HN de. Association Mapping in Plants, Springer New York, 278 p.

Feltus FA, Wasn J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res 14:1812–1819

Fourie H, Mienie CMS, McDonald AH, De Waele D (2008) Identification and validation of genetic markers associated with Meloidogyne incognita race 2 resistance in soybean, Glycine max (L.) Merr. Nematology 10:651–661

Fourie H, Mc Donald AH, De Waele D (2013) Host and yield responses of soybean genotypes resistant or susceptible to Meloidogyne incognita in vivo. Int J Pest Manag 59: 111–121.

Fu, Y.-B. 2012. Genotyping-by-sequencing: A case study in barley. Workshop presented at: Genomics of Genebanks. Plant and Animal Genome Conference XX, San Diego, CA. 14–18 Jan. 2012. Workshop W362.

Ganal M.W.; Altmann T.; Röder M.S. (2009). SNP identification in crop plants. Curr Opin Plant Biol. 12:211-217.

Jammes F, Lecomte P, de Almeida-Engler J, Bitton F, Martin- Magniette M-L, Renou JP, Abad P, Favery B (2005) Genome- wide expression profiling of the host response to root-knot nematode infection in Arabidopsis. Plant J 44:447–458

Jander, G.; Norris, S.R.; Rounsley, S.D.; Bush, D.F.; Levin, I.M.; Last, R.L. (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiol 129:440–450.

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9: 166–177.

Ibrahim H, Hosseini P, Alkharouf N, Hussein E, Gamal El-Din AEK, Aly M, Matthews B (2011) Analysis of gene expression in soybean (Glycine max) roots in response to the root knot nematode Meloidogyne incognita using microarrays and KEGG pathways. BMC Genomics 12:220

Liu X, Liu S, Jamai A, Bendahmane A, Lightfoot D, Mitchum M, Meksem K (2011) Soybean cyst nematode resistance in soybean is independent of the Rhg4 locus LRR-RLK gene. Funct Integr Genomics 11:539–549.

Li, Z., Jakkula, L. & Hussey, R. (2001). SSR mapping and confirmation of the QTL from PI96354 conditioning soybean resistance to southern root-knot nematode. *Theor. Appl. …* 1167–1173.

Liu, S.; Chen, H.D.; Makarevitch, I.; Shirmer, R.; Emrich, S.J.; Dietrich, C.R.; Barbazuk, W.B.; Springer, N.M.; Schnable, P.S. Highthroughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. Genetics 184:19–26, 2010

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, et al. (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Ha B, Bennett JB, Hussey RS, Finnerty SL, Boerma HR (2004) Pedigree Analysis of a Major QTL Conditioning Soybean Resistance to Southern Root-Knot Nematode. Crop Sci Soc 40: 758–763.

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. Theor Appl Genet. Feb;124(3):447-58, 2012

Hayden MJ, Tabone TL, Nguyen TM, Coventry S, Keiper FJ, Fox RL, Chalmers KJ, Mather DE, Eglinton JK (2009) An informa- tive set of SNP markers for molecular characterisation of Australian barley germplasm. Crop Pasture Sci 61:70–83

Hoagland, D.R. and D.I. Arnon. (1950) The water-culture method for growing plants without soil. California Agricultural Experiment Station Circular 347:1-32.

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B,

Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ra ˜tsch G, Buell CR, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci USA 106:12273–12278

Melito S, Heuberger A, Cook D, Diers B, MacGuidwin A, Bent A (2010) A nematode demographics assay in transgenic roots reveals no significant impacts of the Rhg1 locus LRR-Kinase on soybean cyst nematode resistance. BMC Plant Biol 10:104

Milne, I., Bayer, M., Cardle, L., et al., 2010. Tablet – next generation sequence assembly visualization. Bioinformatics 26, 401–402.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, et al. (2012) Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome J 5: 103. Available: https://www.crops.org/publications/tpg/abstracts/5/3/103. Accessed 31 October 2013.

Project G, Asia E, Africa S, Figs S, Tables S (2012) An integrated map of genetic variation. 135: 0–9. doi:10.1038/nature11632.

Pham A-T, McNally K, Abdel-Haleem H, Roger Boerma H, Li Z (2013) Fine mapping and identification of candidate genes controlling the resistance to southern root-knot nematode in PI 96354. Theor Appl Genet 126: 1825–1838. Available: http://www.ncbi.nlm.nih.gov/pubmed/23568221. Accessed 29 January 2014.

Riekert HF, Henshaw GE (1998) Effect of soybean, cowpea and groundnut rotations on root-knot nematode build-up and infestation of dryland maize. Afr Crop Sci J 6:377–383

Sikora RA, Fernández E (2005) Nematode parasites of vegetables. In: Luc M, Sikora RA, Bridge J (eds) Plant parasitic nematodes in subtropical and tropical agriculture, 2nd edn. CABI Publish- ing, Wallingford, pp 319–392

Shearin, Z. P., Finnerty, S. L., Wood, E. D., Hussey, R. S. & Boerma, H. R. (2009). A Southern Root-Knot Nematode Resistance QTL Linked to the -Locus in Soybean. *Crop Sci.* 49, 467.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome, 49, 431–437.

Sonah H, Bastien M, Iquira E, Tardivel A (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8: 1–9. Available: http://dx.plos.org/10.1371/journal.pone.0054603. Accessed 22 December 2013.

Trebbi D, Maccaferri M, de Heer P, Sørensen A, Giuliani S, et al. (2011) High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theor Appl Genet 123: 555–569.

Varshney RK, Spurthi N, Nayak S, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530.

Van Tassel, P.C.; Smith, T.P.L.; Matukumalli, L.K.; Taylor, J.F.; Schnabel, R.D.; Lawley, C.T.; Haudenschild, C.D.; Moore, S.S.; Warren, W.C.; Sonstegard, T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5:247–252.

Xu X, Zeng L, Tao Y, Vuong T, Wan J, et al. (2013) Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. Proc Natl Acad Sci U S A 110: 13469–13474.

Xu X, Zeng L, Tao Y, Vuong T, Wan J, et al. (2013) Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. Proc Natl Acad Sci U S A 110: 13469–13474.

Zhang, X.; Borevitz, J.O. (2009) Global analysis of allele-specific expression in Arabidopsis thaliana. Genetics 182:943–954.

Zhao, W.; Canaran, P.; Jurkuta, R.; Fulton, T.; Glaubitz, J.; Buckler, E.S.; Doebley, J.; Gaut, B.; Goodman, M.; Holland, J.; Kresovich, S.; McMullen, M.; Stein, L.; Ware, D. Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34:D752–D757, 2006.

Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11:267

Yamanaka N, Sato H, Yang Z, Xu DH, Catelli LL, Binneck E, Arias CAA, Abdelnoor RV and Nepomuceno AL (2007) Genetic relationships between Chinese, Japanese, and Brazilian soybean gene pools revealed by simple sequence repeat (SSR) markers. Genet Mol Biol 30:85-88.

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

Wu, X.; Ren, C.; Joshi, T.; Vuong, T.; Xu, D.; Nguyen, H. T. SNP discovery by high-throughput sequencing in soybean. BMC genomics, 11:469-479, 2010

Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–U118.

Wysmierski, P. T. and Vello, N. A. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. *Genet. Mol. Biol.* [online].

4.8 Support information

Supplementary information 01. **List of soybean accessions used in this study**

1	BRS Celeste	50	PI424558A	99	PI458249	148	PI507571
2	BRS282	51	PI424574	100	PI458294	149	PI507602
3	CD201	52	PI424588	101	PI458298	150	PI507609
4	Embrapa 20 (Doko RC)	53	PI424597	102	PI458306A	151	PI508296A
5	PI 59845	54	PI424605A	103	PI458515	152	PI508296G
6	PI054610	55	PI430460A	104	PI458531	153	PI509075
7	PI157428	56	PI430596	105	PI467316	154	PI509074
8	PI157492	57	PI437153A	106	PI470226	155	PI509079
9	PI158765	58	PI437160	107	PI476350B	156	PI518719
10	PI171427	59	PI437341	108	PI483252	157	PI518720
11	PI171431	60	PI437344C	109	PI483253	158	PI520733
12	PI171432	61	PI437350	110	PI495020	159	PI522189
13	PI171454	62	PI437353	111	PI495831	160	PI532455B
14	PI171652	63	PI437423	112	PI495832	161	PI547874
15	PI179826	64	PI437486	113	PI506516	162	PI548493
16	PI196170	65	PI437636B	114	PI506525	163	PI549076A
17	PI200519	66	PI437673	115	PI506590E	164	PI561337
18	PI200538	67	PI437725	116	PI506789	165	PI561346
19	PI229325	68	PI437749	117	PI506819	166	PI561354
20	PI230977	69	PI437773	118	PI506833	167	PI561379B
21	PI248515	70	PI437801	119	PI506848	168	PI567214B
22	PI253651D	71	PI437819	120	PI506862	169	PI567648C
23	PI253652A	72	PI437829	121	PI506892	170	PI567668
24	PI253654	73	PI437845D	122	PI506935	171	PI578397
25	PI253663	74	PI437909A	123	PI506989	172	PI578432A
26	PI304218	75	PI437912	124	PI507072	173	PI578506

27	PI323552	76	PI438048B	125	PI507073	174	PI587608B
28	PI323556	77	PI438123C	126	PI507082A	175	PI587618A
29	PI339868B	78	PI438181B	127	PI507089A	176	PI587991
30	PI371611	79	PI438187	128	PI507089B	177	PI593956A
31	PI374189	80	PI438190	129	PI507097	178	PI593972
32	PI398313	81	PI438193	130	PI507153	179	PI594401B
33	PI417234	82	PI438255	131	PI507158	180	PI594403
34	PI417580	83	PI438302A	132	PI507160	181	PI594427C
35	PI423945	84	PI438303	133	PI507259	182	PI594442B
36	PI424202	85	PI438304B	134	PI507286C	183	PI594470C
37	PI424492	86	PI438307	135	PI507316	184	PI594538A
38	PI424495	87	PI438492	136	PI507317	185	PI594596
39	PI424499D	88	PI442005	137	PI507384	186	PI594775A
40	PI424504A	89	PI442010	138	PI507407	187	PI595099
41	PI424505	90	PI442012A	139	PI507408	188	PI89772
42	PI424506	91	PI442018	140	PI507430	189	PI90490-2
43	PI424511	92	PI442044	141	PI507432	190	PI91178
44	PI424522	93	PI445837	142	PI507443	191	PI96118
45	PI424523B	94	PI458175C	143	PI507447	192	PI96280
46	PI424549A	95	PI458199	144	PI507449	193	PI97038
47	PI424554	96	PI458226	145	PI507480	194	Santa Rosa
48	PI424555B	97	PI458234	146	PI507492		
49	PI424557	98	PI458236A	147	PI507501		

Supplementary information 2. Phenotypic evaluation of **three resistance control and ten resistance accessions showing divergent haplotypes for five SNPs.**

	SNP Position	**1502515**	**1505789**	**1505931**	**1514707**	**1514717**
	Alleles	G/A	C/T	C/A	A/T	G/A
Phenotypic evaluation	Chrom	10	10	10	10	10
1.1	CD201	A	T	A	T	A
1.2	BRS282	A	T	A	T	A
1.0	PI595099	G	C	C	A	G
1.8	PI458298	G	C	C	A	G
1.4	PI438303	G	C	C	A	G
1.7	PI509075	G	C	C	A	G
1.3	PI458234	G	C	C	A	G
1.7	PI437801	G	C	C	A	G
1.7	PI424557	G	C	C	A	G
1.2	PI594775A	G	C	C	A	G
1.8	PI424554	G	C	C	A	G
1.3	PI253654	G	C	C	A	G