Universidade Estadual de Londrina

JOÃO VITOR MALDONADO DOS SANTOS

RESSEQUENCIAMENTO DE GENOMAS DE CULTIVARES BRASILEIRAS DE SOJA:

ANÁLISE ESTRUTURAL E ASSOCIATIVA

Londrina
2015

Universidade Estadual de Londrina

Instituto Agronômico do Paraná

Empresa Brasileira de Pesquisa Agropecuária

JOÃO VITOR MALDONADO DOS SANTOS

RESSEQUENCIAMENTO DE GENOMAS DE CULTIVARES BRASILEIRAS DE SOJA:

ANÁLISE ESTRUTURAL E ASSOCIATIVA

Londrina
2015

Universidade Estadual de Londrina
Centro de Ciências Biológicas
Departamento de Biologia Geral

Programa de
Pós-graduação em
Genética e Biologia Molecular

JOÃO VITOR MALDONADO DOS SANTOS

RESSEQUENCIAMENTO DE GENOMAS DE CULTIVARES BRASILEIRAS DE SOJA:

ANÁLISE ESTRUTURAL E ASSOCIATIVA

Tese apresentada ao Programa de Pós-graduação em Genética e Biologia Molecular da Universidade Estadual de Londrina, como requisito para a obtenção do título de Doutorado.

Orientador: Dr. Ricardo Vilela Abdelnoor
Co-orientadores:
Dr. Henry T Nguyen
Dr. Babu Valliyodan

Londrina
2015

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

Dados Internacionais de Catalogação-na-Publicação (CIP)

S237r Santos, João Vitor Maldonado dos.
Ressequenciamento de genomas de cultivares brasileiras de soja : análise estrutural e associativa / João Vitor Maldonado dos Santos. – Londrina, 2015.
xvi, 113 f. : il.

Orientador: Ricardo Vilela Abdelnoor.
Coorientadores: Henry T. Nguyen, Babu Valliyodan.
Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2015.
Inclui bibliografia.

1. Soja – Resistência a doenças e pragas – Aspectos genéticos – Teses. 2. Seqüência de nucleotídeos – Teses. 3. Soja – Melhoramento genético – Teses. 4. Plantas – Filogenia – Teses. I. Abdelnoor, Ricardo Vilela. II. Nguyen, Henry T. III. Valliyodan, Babu. IV. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. V. Instituto Agronômico do Paraná. VI. EMBRAPA. VII.Título.

CDU: 631.52:633.34

JOÃO VITOR MALDONADO DOS SANTOS

RESSEQUENCIAMENTO DE GENOMAS DE CULTIVARES BRASILEIRAS DE SOJA:

ANÁLISE ESTRUTURAL E ASSOCIATIVA

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

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Ricardo Vilela Abdelnoor
Empresa Brasileira de Pesquisa Agropecuária - Embrapa/Soja

_____

Prof. Dr. Everaldo Gonçalves de Barros
Universidade Católica de Brasília - UCB

_____

Profª Dra. Mayra Costa da Cruz Gallo de Carvalho
Universidade Estadual do Norte do Paraná - UENP

_____

Profª Dra. Francismar Corrêa Marcelino-Guimarães
Empresa Brasileira de Pesquisa Agropecuária - Embrapa/Soja

_____

Prof. Dr. Douglas Silva Domingues
Universidade Estadual de Londrina - UEL

Londrina, 05 de Março de 2015.

DEDICO

Ao meu avô Manoel (in memoriam) por
me guiar a escolher este caminho.

OFEREÇO

Aos meus pais por todo apoio, paciência e dedicação em busca dos meus objetivos.

AGRADECIMENTOS

A Deus que me iluminou nos momentos difíceis e sempre esteve presente em minha vida.

As minhas duas irmãs, Aliny e Rubia, e meus dois sobrinhos, Davi e Miguel, pelo amor, incentivo e apoio incondicional.

A Universidade Estadual de Londrina, ao programa de pós-graduação em Genética e Biologia Molecular e seu corpo docente pela ótima formação que me proporcionaram.

A coordenadora do curso Dra. Ana Lúcia Dias e a ex-secretária Sueli Trindade Miranda por todo auxílio prestado, carinho e respeito durante essa caminhada.

Ao Centro Nacional de Pesquisa em Soja (Embrapa Soja) e a Universidade de Missouri (MIZZOU), pela oportunidade e suporte financeiro cedido durante meu doutorado.

A CAPES pelo suporte financeiro e pela bolsa concedida tanto no Brasil quanto no período sanduíche durante meu doutorado.

Ao meu orientador Dr. Ricardo Vilela Abdelnoor por todo auxílio intelectual, atenção e grande contribuição na realização do trabalho.

Aos meus co-orientadores, Dr. Henry Nguyen e Dr. Babu Valliyodan pelas contribuições intelectuais, auxílio no trabalho e recepção nos Estados Unidos.

Aos pesquisadores da Embrapa Soja, Dr. Carlos Alberto Arrabal Arias, Dra. Francismar Corrêa Marcelino-Guimarães, Dr. Rafael Moreira Soares e Dr. Waldir Pereira Dias o por todo o auxílio e sugestões dadas ao trabalho.

Aos pesquisadores da Universidade de Missouri, Dr. Dong Xu, Dr. Silvas Prince e Dr. Tri Vuong, pelas discussões construtivas e contribuição intelectual ao trabalho.

Aos membros do laboratório de biologia digital da Universidade de Missouri, em especial a pesquisadora Dra. Joshi Trupti, por todo apoio técnico, intelectual e aprendizado conquistado.

Aos membros do laboratório de genética molecular e genômica da soja da Universidade de Missouri, em especial a Jordan Mroz e Theresa Musket, pela amizade e todo suporto para realização do trabalho.

Aos atuais e ex-funcionários do laboratório de Biotecnologia Vegetal da Embrapa Soja, César, Danielle, Márcia, Jairo, Silvana, Renan e Verinha por todo auxílio e amizade.

Aos atuais e ex-bolsistas do laboratório de Biotecnologia Vegetal e Bioinformática da Embrapa Soja, pelo companheirismo e amizade durante estes quatro anos de doutorado.

A Aguida por todo auxílio, suporte e conselhos dados durante o desenvolvimento da tese.

A Valéria por todas as sugestões para elaboração da tese e por sempre estar presente nos momentos bons e ruins destes quatro anos

Ao pessoal da turma de doutorado em Genética e Biologia Molecular, pelo ótimo convívio durante e as amizades formadas durante estes dois anos que levarei para sempre, em especial a Fabiola, Juliana e Michelle.

A todos os amigos que obtive no período sanduíche realizado nos Estados Unidos, em especial a Guilherme (Gaúcho), Jerry, Júnior, Renan, Samuel (Bahia), Tadeu e Thiago, pelo convívio de um ano e amizade que levarei por toda uma vida.

Aos meus “irmãos” americanos Mike Jones, Blake Boland e Codey Scott por todo convívio e bons momentos vividos no meu período sanduíche.

Aos meus grandes amigos de Londrina, Alan, Daniel, Guilherme, Leandro

(Shampoo), Lucas e Paolo, pela grande força e amizade em todos os momentos.

Aos meus amigos de Tupã, Arthiê, Adão, Danilo, Fred, Eduardo Goldoni, Joel, Leandro Castro, Lucas Castro, Lucas Leal, Serginho, Thiago Maciel, Thiago Sato, Diego (Jorge), Fernando, Marcel e Tiago Franco pela amizade e por estarem ao meu lado nos momentos bons e ruins.

Aos amigos Adriano, Beca, Bruno, Cesar, Davi, Fernando Ferric, Fernando Tochete, Guilherme, Gustavo Ibraim, Gustavo Scatolin, Israel, João Paulo, Júlio, Rone, Tiago e meu primo, Luís Gustavo, por todo apoio e amizade cultivada ao longo do tempo.

A todos que de alguma forma contribuíram diretamente ou indiretamente para minha formação e desenvolvimento deste trabalho.

SANTOS, João Vitor Maldonado dos. **Ressequenciamento de genomas de cultivares brasileiras de soja**: análise estrutural e associativa. 2015. 113f. Tese (Doutor em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina, 2015.

RESUMO

A soja [*Glycine max* (L.) Merrill] é uma das mais importantes leguminosas cultivadas no mundo devido sua importância na alimentação humana, animal e produção de biocombustíveis. Desde que seu genoma foi sequenciado, aumentou-se o interesse por estudos de variações alélicas e estruturais ligadas a importantes características agronômicas e associadas à resistência a patógenos. Neste trabalho, foram ressequenciados 28 cultivares brasileiras com o objetivo de investigar sua base genética e análise de associação ampla do genoma para identificar importantes variações alélicas ligadas a resistência contra nematoide de cisto, nematoide de galha e cancro da haste. Foram identificados 1.329.844 *indels* e 5.868.344 SNPs, sendo 10.079 SNPs relacionados à resistência as três doenças. Uma estrutura homogênea foi observada na base genética nacional, com a presença de possíveis regiões sob processo de seleção positiva. Ainda, regiões contendo CNVs foram identificadas no genoma da soja. Os resultados indicam que apesar da base genética nacional estar se diversificando, ainda continua estreita, o que aumenta a necessidade de utilização de acessos de outras localidades para aumentar a variabilidade da base genética da soja brasileira. As variações alélicas relacionadas à resistência a doenças possuem aplicação direta para programas de melhoramento, devendo ser validadas futuramente. Por fim, os CNVs detectados podem contribuir significativamente no aumento da produtividade, e ainda estar relacionados à seleção de combinações que levam a uma resistência maior dos genótipos analisados contra as doenças estudadas. Uma análise mais aprofundada de tais variações estruturais torna-se importante em futuros trabalhos.

Palavras-chave: *Glycine max*. Variação alélica. GWAS. Diversidade genética. Seleção positiva desequilíbrio de ligação. CNVs.

SANTOS, João Vitor Maldonado dos. **Reassessment of genomes of Brazilian soybean cultivars:** structural and associative analysis. 2015. 113f. Thesis (PhD in Genetics and Molecular Biology) – Universidade Estadual de Londrina, Londrina, 2015.

ABSTRACT

Worldwide, soybean [*Glycine max* (L) Merrill] is one of the most important crops due to the major importance in human food, animal feed and biofuel production. Since its reference genome was sequenced, interest has grown on structural and allelic variation that can be related to important agronomical traits and resistance against pathogens. In this study, we re-sequenced 28 Brazilian soybean cultivars to investigate the genetic base and a GWA analysis to identify important allelic variation related to resistance against soybean cyst nematode, root-knot nematode and stem canker. We identified 1,329,844 indels and 5,868,344 SNPs, being 10,079 SNPs related to resistance mechanisms against the three diseases analyzed in this work. A homogeneous structure can be observed in the Brazilian genetic basis, with the presence of possible regions under positive selection processes. Additionally, CNVs regions were also detected on soybean genome. The results suggested that despite the fact that Brazilian soybeans are diversifying, the genetic base is still narrow, which underline the need to introduce new exotic alleles from other germplasm in Brazilian genetic basis. Furthermore, allelic variations related to resistance against diseases have directly application for breeding programs and should be validated. Finally, the CNVs detected in this work could play a key role for increase the Brazilian soybean production, and may also be related to resistance mechanisms against the analyzed diseases.

Keywords: *Glycine max*. Allelic variation. GWAS. Genetic diversity. Purifying selection. Linkage disequilibrium. CNVs.

LISTA DE FIGURAS

LISTA DE TABELAS

ARTIGO I – EVALUATION OF GENETIC VARIATION BY GENOME RESEQUENCING OF 28 BRAZILIAN SOYBEAN CULTIVARS

ARTIGO II - RESEQUENCING STRATEGIES FOR GENOME-WIDE ASSOCIATION STUDIES FOR IMPORTANT BIOTIC STRESSES IN SOYBEAN

SUMÁRIO

1. INTRODUÇÃO

Mundialmente, a agricultura encontra-se como uma das principais fontes de desenvolvimento socioeconômico. As constantes mudanças climáticas associadas ao crescente aumento populacional e ao impacto econômico causado pela agricultura em um país tornam essenciais as pesquisas relacionadas a grandes culturas. Assim, o desenvolvimento de tecnologias e produtos que visam um aumento significativo na produção e qualidade mundial das principais culturas torna-se extremamente importante. Dentro deste panorama, a soja [*Glycine max* (L) Merrill] surge como uma das principais *commodities* comerciais que movimenta o mercado internacional. No Brasil, possui um papel extremamente importante na economia, sendo o principal produto de exportação do agronegócio. Desta forma, estudos relacionados a esta leguminosa são cada vez mais importantes para manutenção do Brasil entre os maiores produtores mundiais.

Esta posição brasileira de destaque no cenário mundial somente foi alcançada com auxilio dos programas de melhoramento genético da cultura, que embora recente quando comparado a outras culturas, já trouxeram grandes avanços. Apesar das melhorias obtidas, os programas de melhoramento genético da soja ainda precisam contornar alguns problemas, tais como os fatores abióticos e bióticos que causam prejuízos significativos à produção nacional. Além disto, pela história recente da cultura no país e pela utilização de poucos ancestrais oriundos da base genética americana, espera-se que a base genética da soja nacional possua pouca variabilidade, o que dificulta melhoristas a produzirem cultivares com desempenho cada vez melhor em campo e com maior qualidade.

Assim, ferramentas que possam auxiliar programas brasileiros de melhoramento da soja são fundamentais a sobrevivência do agronegócio brasileiro e da economia nacional como um todo. Dentre elas, técnicas de biologia molecular surgem como importantes ferramentas que permitem a identificação de genes que possam auxiliar na melhoria da qualidade e produtividade da cultura, o que auxilia programas de melhoramento a acelerar o processo de produção de cultivares-elites preparadas para um mercado extremamente competitivo. Com o sequenciamento do genoma da soja, uma grande quantidade de informação encontra-se disponível em banco de dados públicos e várias ferramentas de bioinformática já foram desenvolvidas com aplicabilidade direta ao melhoramento de plantas. Neste

contexto, com a grande quantidade de informação disponível e acessível, estudos de resequenciamento sobre as principais cultivares nacionais, assim como de acessos oriundos de outras localidades que contenham características específicas importantes, como tolerância a estresses bióticos e abióticos, surgem como alternativas na busca de variações alélicas que possam auxiliar no aumento da variabilidade genética do germoplasma brasileiro e, consequentemente expressivo impacto na produtividade, fitossanidade e qualidade das cultivares nacionais de soja.

2. OBJETIVOS

2.1. Objetivo geral

O presente trabalho possui como objetivo:

- Identificar importantes variações alélicas e estruturais presentes nos acessos comerciais da soja brasileira que possam estar relacionadas a importantes características agronômicas, como doenças.

2.2. Objetivos específicos

- Detectar potenciais SNPs e *indels* no genoma dos acessos resequenciados;
- Encontrar variações alélicas ligadas à resistência genética contra nematoide de cisto, nematoide de galhas e cancro da haste;
- Identificar a ocorrência de queda de desequilíbrio de ligação entre os acessos resistentes analisados;
- Compreender a relação filogenética entre os acessos estudados;
- Avaliar a estrutura populacional e diversidade dos acessos estudados;
- Comparar a base genética da soja brasileira com a norte-americana;
- Analisar as modificações estruturais presentes no genoma dos acessos sequenciados.

3. REVISÃO BIBLIOGRÁFICA

3.1. Soja

A soja pertence à divisão Magnoliophyta, classe Magnoliopsida, subclasse Rosidae, ordem Fabales, família Fabaceae, subfamília Faboideae, gênero *Glycine* L., subgênero Glycine subg. soja (Moench) e espécie *Glycine max* (L.) Merril. É considerada a leguminosa mais importante do mundo, devido principalmente a sua importância na alimentação humana e animal e produção de biocombustíveis. Hoje, a mesma possui um lugar de destaque na economia brasileira, sendo o maior produto nacional de exportação ((EMBRAPA SOJA, 2014).

Trata-se de uma espécie autógama, com baixo índice de polinização cruzada devido à cleistogamia. Contudo, ainda existe uma taxa de fecundação cruzada de 1% (BORÉM; MIRANDA, 2005), causada devido a ação de agentes polinizadores, principalmente insetos como abelhas e tripses (SEDIYAMA et al., 1986). É estimado que seu ancestral selvagem (*Gycine soja*) tenha sido domesticado por volta de 7.000 – 9.000 anos atrás, na Ásia (LEE et al., 2011). No Brasil, a cultura ganhou destaque no final da década de 60, associada à expansão da triticultura no Rio Grande do Sul, sendo usada como opção no verão para rotação de cultura. Além disto, o farelo de soja era importante material utilizado para alimentação de suínos e aves (EMBRAPA SOJA, 2014).

A grande expansão da cultura no país aconteceu em meados da década de 70 com a explosão do preço da soja no mercado mundial. O Brasil possui uma vantagem competitiva comparada aos demais países produtores, pois o escoamento de sua safra ocorre no período de entressafra dos Estados Unidos. Com isto, investimentos em tecnologias para adaptação a outras regiões do país foram realizados, o que levou a “tropicalização” da soja. Assim, tornou-se possível a utilização do grão em regiões de baixas latitudes, o que causou uma revolução na produção da cultura, sendo seu impacto constatado pelo mercado a partir do final da década de 80 e mais notoriamente na década de 90, com a queda no preço do grão (EMBRAPA SOJA, 2014).

Globalmente, os líderes mundiais em produção de soja são Estados Unidos, Brasil, Argentina, China, Índia e Paraguai (EMBRAPA SOJA, 2014). Estima-se que na safra 2014/15, o Brasil produzirá uma safra estimada de 95.919.800 toneladas, plantadas em 31.162.180 hectares, com uma média de produtividade de 3.033 kg/ha. Os maiores produtores nacionais são os estados do Mato Grosso, com

estimativa de 28.216.400 toneladas, seguida pelo Paraná, com 17.224.700 toneladas (COMPANHIA NACIONAL DE ABASTECIMENTO, 2014).

3.2. Estratégias de sequenciamento aplicadas ao genoma da soja

O elevado número de plataformas de sequenciamento permitiu gerar uma grande quantidade de trabalhos com genomas completos. Em soja, existem vários trabalhos utilizando tais estratégias, sendo a cultivar Williams 82 a primeira sequenciada por completo. Foram gerados 1,1 gigabase do genoma sequenciado via *Whole Genome Shotgun* e preditos aproximadamente 46.430 genes codificadores de proteínas distribuídos nos 20 cromossomos da soja. Estes valores são cerca de 70% maiores que os observados em *Arabidopsis thaliana*. Além disso, os autores verificaram que aproximadamente 75% dos genes presentes no genoma da soja estão de múltiplas cópias (SCHMUTZ et al., 2010).

O sequenciamento completo do genoma de acessos de soja selvagem (*Glycine soja*) também já foram realizados. Kim et al. (2010) sequenciaram 915,5 megabase do genoma de um acesso de soja selvagem e observaram a existência de 2,5 megabase de sequências substituídas, 406 kilobase de pequenas deleções/inserções, 32,4 megabase de grandes deleções e 8,3 megabase de novas sequencias quando comparado com o genoma referência de *Glycine max*.

Um levantamento realizado mostrou a existência de mais de 120.000 sequencias de nucleotídeos, 1.460.000 ESTs, 368.000 sequencias genômicas, 80.000 sequências de proteínas e mais de seis milhões de SNPs oriundos da análise do genoma da soja depositados no banco de dados do GenBank (NCBI) (BENKO-ISEPPON; NEPOMUCENO; ABDELNOOR, 2012). O grande número de informações depositadas constantemente em banco de dados disponíveis na internet demonstra a importância de estudos para melhor compreender as informações genéticas desta leguminosa Em meio à grande quantidade de informações geradas pelo sequenciamento completo de genomas, estratégias de resequenciamento surgem como importante ferramenta para estudos de variações alélicas. Um estudo comparativo identificou a presença de regiões com alta diversidade entre 31 acessos de soja selvagens e comerciais. Além disto, foram encontrados 205.614 SNPs que podem ser utilizados em estudos associativos e programas de mapeamento de QTLs (LAM et al., 2010). Outro trabalho identificou

um total de 3,87 milhões de SNPs de alta qualidade entre acessos comerciais e selvagens coreanos (CHUNG et al., 2014). Já Li e colaboradores (2013) analisaram por resequenciamento 25 novos acessos de sojas chinesas e 30 acessos de sojas localizadas em banco de dados públicos e identificaram um total de 5.102.244 SNPs e 707.969 inserções/deleções, das quais 25,5% não foram descritas em outros trabalhos. Recentemente, Zhou e colaboradores (2015) resequenciaram 62 acessos de soja selvagem, 130 acessos primitivos e 110 acessos comerciais modernos de soja e encontraram 230 regiões sofrendo influência de processos seletivos, sendo 96 relacionados com QTLs para óleo e 21 contendo genes para biossíntese de ácidos graxos. Além disto, o mesmo trabalho observou alguns SNPs associados com a distribuição geográfica dos acessos de soja.

A presença de uma grande quantidade de trabalhos de sequenciamento completo do genoma em soja, associado a grande quantidade de informação disponível em bancos de dados atua como uma importante ferramenta para programas de melhoramento genético da soja. Assim, abre-se a possibilidade de aplicação direta de resultados de sequenciamento principalmente para busca por genes, QTLs e variantes alélicas de interesse agronômico, ao desenvolvimento de marcadores para seleção assistida, à clonagem baseada em mapeamento e estudos de genética de população e de associação (CARVALHO; SILVA, 2010; LAM et al., 2010; ZHOU et al., 2015).

3.3. Estudos de associação ampla do genoma

O grande objetivo de um programa de melhoramento genético é a identificação e seleção de combinações adequadas de alelos controladores de características de interesse para serem aplicados em processos de seleção artificial. Estas variações alélicas localizam-se em regiões não expressas do genoma ou regiões codantes e reguladoras da expressão de genes, sendo estas duas últimas, mesmo em menor proporção no genoma, componentes diretos da variação fenotípica observada em coleções de germoplasma ou populações naturais (FERREIRA; GRATTAPAGLIA, 2006).

Dentre as estratégias de detecção de variações alélicas no genoma, encontram-se os estudos de associação ampla do genoma (Genome-Wide Association Studies – GWAS). De acordo com o instituto nacional de saúde dos Estados Unidos, trata-se de qualquer estudo de variação genética comum em todo o

genoma humano concebido para identificar associações genéticas com características observáveis (NIH). O primeiro trabalho relacionado a esta estratégia foi observado em humano, para degeneração macular relacionada à idade (KLEIN, 2005). Uma vasta quantidade de trabalhos na literatura já foi descrito para aplicação desta metodologia em humanos. Embora exista essa grande quantidade de trabalhos voltados para área humana, esta estratégia vem sendo amplamente empregada em estudos de genética de plantas. Aranzana et al. (2005) identificaram em *Arabidopsis thaliana* genes previamente conhecidos para período de floração e resistência a patógenos (*Rpm1*, *Rps2* e *Rps5*) em 95 acessos que possuíam dados de polimorfismo de genoma completo disponíveis. Existem inúmeros outros estudos relacionados com outros organismos, como cevada (MASSMAN et al., 2010), arroz (HUANG et al., 2010), sorgo (MORRIS et al., 2013), pimenta (QIN et al., 2014) e milho (KUMP et al., 2011).

Um grande número de trabalhos já foi descrito na literatura sobre GWAS aplicados ao genoma da soja. Hwang et al. (2014) identificaram 40 SNPs em 17 diferentes regiões genômicas associados com teor de proteínas na semente. Além disto, 25 SNPs em 13 regiões genômicas diferentes foram identificados relacionados com regiões controladoras do teor de óleo da semente. Outro trabalho realizado com soja detectou a presença de três QTLs nos cromossomos 3, 8 e 20 com associação significativa de resistência contra *Sclerotinia sclerotiorum* (IQUIRA; HUMIRA; FRANÇOIS, 2015). Hao et al. (2012) identificaram 51 SNPs relacionados com teor e fluorescência da clorofila, sendo 14 destes SNPs coassociados com características de produtividade. Por fim, Mamidi et al. (2011, 2014) realizaram trabalhos na detecção de SNPs relacionados com a clorose causada pela deficiência de ferro. No primeiro trabalho, duas populações 2005 [(n = 143) e 2006 (n = 141)] foram avaliadas. Um total de 42 loci para 2005 e 88 para 2006 foram estatisticamente associados à clorose causada pela deficiência de ferro, sendo 9 presente em ambas as populações. Já no Segundo trabalho, um total de 20 genes-candidatos conhecido por estarem relacionados ao metabolismo de ferro foram associados a regiões de QTLs. Ainda, foi identificado um SNP não sinônimos um QTL no cromossomo 3 sugerindo que o ortólogo do FRE1 é o maior responsável pela variação fenotípica observada na soja. O surgimento e avanço das novas tecnologias para sequenciamento em larga escala dos genomas possibilitou a

identificação de um grande número de variações alélicas com aplicabilidade em GWAS em plantas.

3.4. Variações no número de cópias – Copy-number variations

As variações denominadas de variações no número de cópias (Copy-number variations - CNVs) referem-se a modificações estruturais que produzem alterações no número de cópias numa região específica do genoma. Tais modificações podem variar em tamanho e hoje são amplamente estudadas. Três mecanismos surgem como possíveis principais fontes de formação de regiões contendo CNVs no genoma: recombinação homóloga não alélica (non-allelic homologous recombination - NAHR), modelo de troca de forquilha de replicação (Fork Stalling and Template Switching - FoSTeS) e união terminal não homóloga (non-homologous end-joining - NHEJ).

Recombinações homólogas não alélicas geralmente ocorrem em hotspots entre segmentos de DNA de elevada similaridade, mas que não são alelos e comumente envolve segmentos de DNA de repetições com baixo número de cópia (Low Copies Repeats - LCRs). Podem ser de três tipos dependendo de sua localização (rearranjos intracromátides, intercromátides ou intercromossomal) e orientação (direta, oposta ou mista), podendo ocorrer em células meióticas ou mitóticas (GU; ZHANG; LUPSKI, 2008; ZMIEŃKO et al., 2014). Deleções e duplicações podem surgir quando duas LCRs encontram-se no mesmo cromossomo e na mesma orientação (direta). Inversões surgem com LCRs no mesmo cromossomo, mas em orientação oposta. Por fim, LCRs localizadas em diferentes cromossomos podem originar translocações cromossomais (GU; ZHANG; LUPSKI, 2008) (**Figura 7**).

Outro mecanismo potencial para surgimento de regiões de CNVs são os FoSTeS, causados por erros durante o processo de replicação do DNA. Ocorre após uma fita de DNA que está sendo sintetizada em uma forquilha de replicação se desacopla e invade uma nova forquilha de replicação fisicamente próxima, onde se anelará e continuará a síntese de DNA de maneira errada (LEE; CARVALHO; LUPSKI, 2007). A mudança entre moldes de DNA é dirigida pela presença de microhomologias entre a fita de DNA molde original e a invadida (**Figura 8b**). Podem gerar deleções, inserções e outros arranjos mais complexos dependendo da quantidade de eventos de mudanças ou o lugar em que a forquilha foi invadida

(upstream ou downstream a forquilha usada inicialmente) (GU; ZHANG; LUPSKI, 2008; ZMIEŃKO et al., 2014)

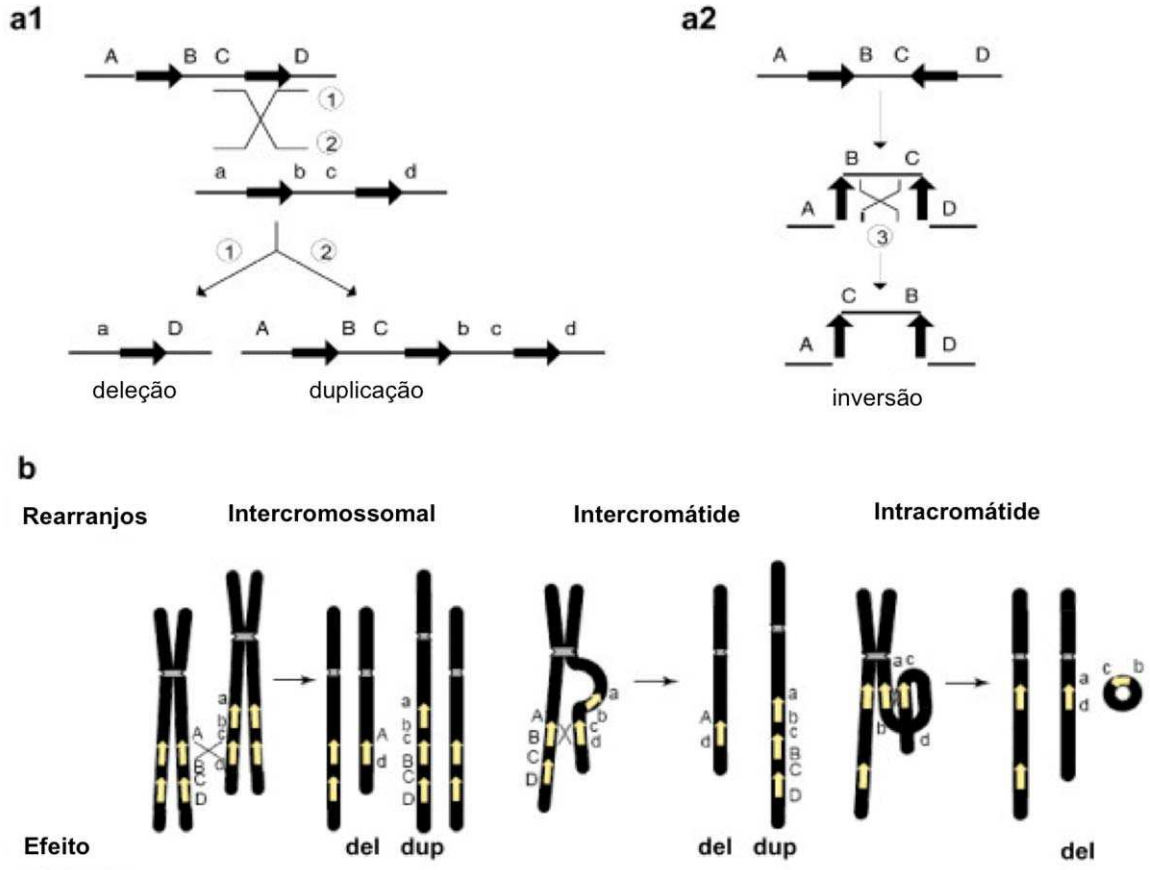


Figura 7. Rearranjos genômicos resultantes da recombinação em regiões contendo LCRs, de acordo com sua orientação (a) e localização (**b**). Figura adaptada de Gu, Zhang e Lupski (2008).

Por fim, NHEJ é um dos mecanismos mais frequentes para reparo de quebras de dupla fita em células eucarióticas. Em humanos, observa-se a utilização deste mecanismo para reparar quebras de dupla fita patológicas, “fisiológicas” ou recombinações, causadas por formas reativas de oxigênio ou radiação ionizante. É considerado o principal responsável por translocações cromossômicas em células cancerosas. Ocorre em quatro etapas: primeiramente ocorre a detecção da quebra da dupla fita, seguido da realização de uma ponte molecular entre as terminações da quebra, modificação das terminações para que aconteça a compatibilização e/ligação e por fim, a ligação das fitas de DNA (GU; ZHANG; LUPSKI, 2008; WETERINGS; VAN GENT, 2004) (**Figura 8a**). Ainda, este processo possui duas importantes características: não é dependente de regiões LCRs ou de segmentos de

processamento mínimo eficiente (minimal efficient processing segments - MEPS), mas podem estimular o processo, como foi observado por Stankiewicz et al., 2003. Além disto, tal processo deixa uma “cicatriz” no sítio de recombinação devido a inserção ou remoção de vários nucleotídeos no começo ou no término das extremidades (GU; ZHANG; LUPSKI, 2008; LIEBER, 2008). Assim quebras consecutivas e junções indevidas resultam em modificações significativas na estrutura genômica (GU; ZHANG; LUPSKI, 2008).

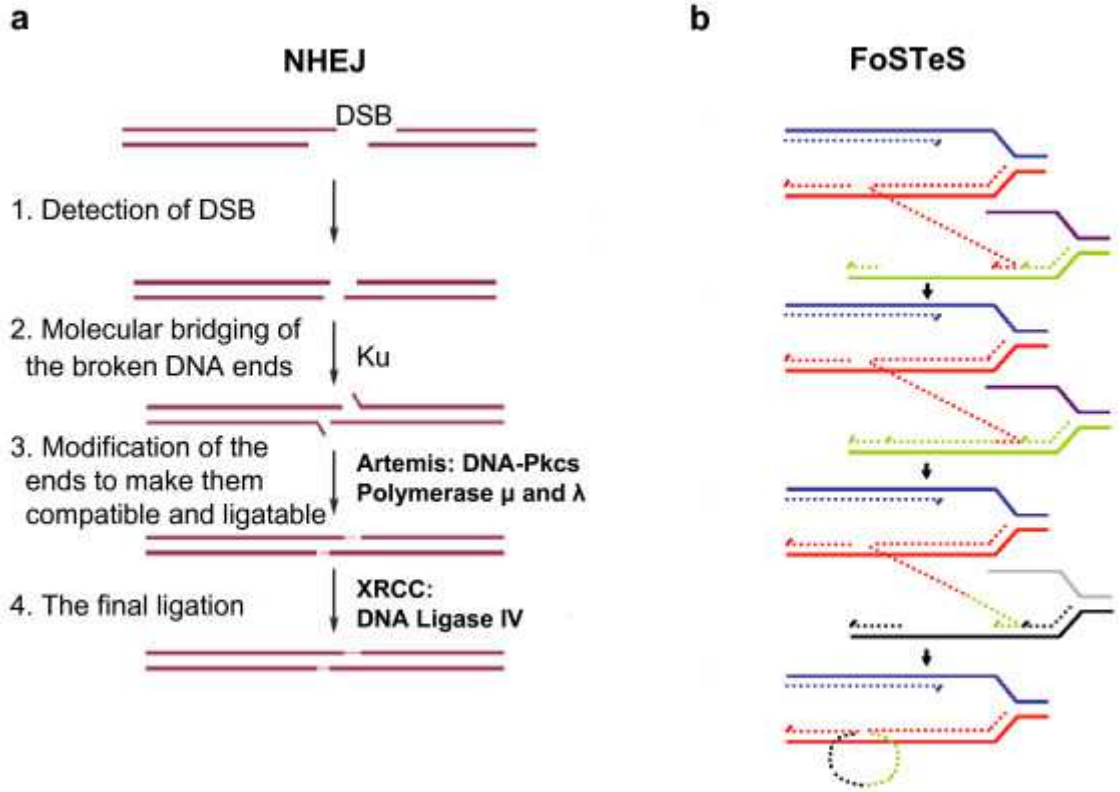


Figura 8. Demais mecanismos responsáveis pelo surgimento de variações estruturais no genoma. Mecanismos NHEJ (a) e FoSTeS (GU; ZHANG; LUPSKI, 2008).

Em humanos, trabalhos apontam que tais modificações estruturais são comuns no genoma (IAFRATE et al., 2004; SEBAT et al., 2004; TUZUN et al., 2005), podendo causar mudanças fenotípicas drásticas, por interromper sequencias codificantes ou perturbando regiões reguladoras do gene (KLEINJAN; VAN HEYNINGEN, 2005). Inúmeras desordens humanas já foram diagnosticas associadas a CNVs, tais como Alzheimer (ROVELET-LECRUX et al., 2006), Parkinson (SIMON-SANCHEZ et al., 2008) e autismo (SEBAT et al., 2010).

Atualmente, observa-se também a presença de estudos em busca de CNVs em uma grande variedade de culturas. Contudo, ainda apresenta-se pouco explorado para análise de tais regiões no genoma de plantas. São descritas análises

de CNVs em *Arabidopsis thaliana* (CAO et al., 2011; DEBOLT, 2010; LU et al., 2012), arroz (YU et al., 2011), cevada (MUÑOZ-AMATRIAÍN et al., 2013), sorgo (ZHENG et al., 2011) e milho (LAI et al., 2010; SPRINGER et al., 2009).

Em soja, alguns trabalhos são descritos com a presença de regiões contendo CNVs. Em um trabalho de resequenciamento, identificou-se 5.500 variações de presença ou ausência em comparação com um acesso selvagem e a cultivar referencia Williams 82 (LAM et al., 2010). Recentemente, outro trabalho de resequenciamento encontrou 162 regiões de CNVs potencialmente selecionadas durante processos de domesticação e melhoramento (ZHOU et al., 2015). Além de avaliações gerais do genoma, ainda foram observadas ligações diretas de CNVs a genes de resistência a estresses bióticos da soja. Para nematoide de cisto, observou-se uma correlação direta na atuação do gene *Rhg1* como a resistência a nematoide de cisto, já que acessos considerados suscetíveis possuíam uma cópia em tandem do cluster gênico, enquanto os resistentes apresentavam 10 cópias (COOK et al., 2012). Outro trabalho demonstrou que a aptidão de avirulência de *Phytophthora sojae* sobre a soja está relacionada com a presença de variações do número de cópias *Avr1a* e *Avr3a* do patógeno (QUTOB et al., 2009). Por fim, foram encontrados CNVs nos cromossomos 3, 6, 7, 16 e 18 em regiões ricas em classes de receptores de proteínas quinases e ligação de nucleotídeos, sendo ambas muito importantes em mecanismos de defesa da planta (MCHALE et al., 2012). Portanto, a identificação de tais CNVs no genoma da soja torna-se extremamente importante e pode servir como chave para compreensão das principais modificações estruturais de seu genoma, bem como identificar potenciais genes-alvo a serem mais bem estudados funcionalmente.

4. MATERIAL E MÉTODOS

4.1. Material genético e sequenciamento

Vinte e oito cultivares de sojas brasileiras foram selecionadas para este trabalho. Tais cultivares foram selecionadas baseadas em sua data de comercialização e grupo de maturidade relativo. As sementes foram obtidas do banco de germoplasma da Embrapa Soja. Amostras de tecido foliar jovem de cada uma das 28 cultivares brasileiras foram coletadas no estádio V_3 da soja. O DNA genômico das cultivares coletadas foram extraídos através do kit de extração Qiagen Mini Plant DNeasy (Qiagen Inc., Valência, CA, EUA), seguindo as instruções do kit. O sequenciamento do material foi realizado na empresa FASTERIS, Suíça, através da plataforma Illumina Hiseq 2000, o que gerou 100 pb de reads paired-end, com uma cobertura esperada de 15 vezes o genoma da soja. Amostras resequenciadas de acessos oriundos dos Estados Unidos e da Ásia foram gentilmente providenciados pelo laboratório de genética molecular e genômica da soja da Universidade de Missouri. Os acessos americanos e asiáticos foram nomeados de acordo com um código devido a privacidade dos dados. Tais genomas foram utilizados para análises de diversidade, GWA e comparações de CNVs.

4.2. Informação fenotípica

Os dados fenotípicos utilizados neste trabalho foram gentilmente cedidos pela Embrapa soja e pelo laboratório de genética molecular e genômica da soja da Universidade de Missouri. Os dados fenotípicos para resistência de todos os acessos contra cancro da haste foram fornecidos pelo laboratório de fitopatologia da Embrapa Soja, sendo os acessos classificados em resistentes, moderadamente resistentes e suscetíveis. Placa contendo meio de cultura BDA foram preparadas contendo cerca de 50 a 150 palitos de dentes. Após a preparação de tais placas, cinco a seis discos do fungo *Diaporthe phaseolorum var. meridionalis* foram inseridos por placas. Estas foram colocadas em câmara de incubação a 25ºC e cinco a seis dias depois, encontravam-se prontas para serem inoculadas. Todos os genótipos que seriam testados contra o fungo foram semeados em vasos de terra. Um total de 15 sementes por vaso foi utilizado para representar cada acesso usado. Adicionalmente, seis vasos contendo sementes da cultivar BR 23 foram adicionados como testemunhas. A inoculação ocorreu no estádio V3 da planta, espetando-se os

palitos de dentes colonizados cerca de um centímetro abaixo do nódulo cotiledonar. A avaliação para cada acesso ocorreu 20 dias depois, com uma nota sendo determinada de acordo com o número de plantas saudáveis restantes por vaso. Vasos contendo mais de 50% de plantas saudáveis eram considerados acessos resistentes, enquanto que valores abaixo de 50% eram acessos suscetíveis.

Adicionalmente, o laboratório de nematologia da Embrapa Soja gentilmente providenciou informações de resistência das cultivares nacionais contra nematoides de galha e cisto. Dados fenotípicos de resistência de cada cultivar brasileira foram gerados em estudos anteriores (Waldir Pereira Dias, informação pessoal). Em contrapartida, seis plantas para cada genótipo foram avaliadas para análise de resistência contra RKN em casa-de-vegetação. Os acessos foram semeados em tubos plásticos contendo 500 cm3 de uma mistura de areia e terra autoclavada (3:1). As plântulas foram mantidas durante 16 horas de iluminação e posteriormente complementadas com lâmpadas de sódio de alta pressão (sistema de luz PL). Uma suspensão de 5.000 ovos de M. incógnita foi inoculada em cada tubo no estádio V2 da planta. Foi utilizada uma solução nutritiva de 80 ml (HOAGLAND; ARNON, 1950) semanalmente como forma de fertilizar as plantas. Trinta dias após a inoculação, raízes de cada planta foram avaliadas de acordo com uma escala que ia de um a cinco baseadas na abundância de galhas. A resposta dos acessos contra RKN foi medida de acordo com a resposta média das seis plantas por genótipo. Cultivares resistentes possuíam notas entre 1,0 – 3,0, enquanto que as suscetíveis tinham notas entre 3,1 – 5,0.

Os dados de resistência dos acessos americanos e asiáticos contra nematoides de galha e cisto foram gentilmente cedidos pelo laboratório de genética molecular e genômica da soja da Universidade de Missouri. As avaliações fenotípicas de nematoide de cisto separaram os acessos em resistentes, moderadamente resistentes, moderadamente suscetíveis e suscetíveis, enquanto as avaliações de nematoide de galhas agruparam os acessos em resistentes ou suscetíveis.

4.3. Detecção de SNPs e Indels

As reads geradas pelo resequenciamento dos acessos nacionais foram mapeadas na nova versão do genoma referencia da soja (Gmax_275_Wm82.a2.v1) através do programa de alinhamento *Burrows-Wheeler Aligner* (BWA) (LI; DURBIN,

2009). Após o mapeamento, as reads alinhadas foram processadas pela ferramenta *Piccard* versão 1.107 com a finalidade de remover valores duplicados. Um arquivo binário de extensão ".bam" foi gerado, representando o genoma montado de cada cultivar resequenciada. Para busca de SNPs/Indels, foi utilizado o conjunto de ferramentas *Genome Analysis Toolkit* (GATK) versão 3.0 (MCKENNA et al., 2010). Este conjunto de ferramentas foi utilizado para realização de um realinhamento em regiões de Indels e uma recalibração qualitativa com a finalidade de gerar um novo arquivo binário de extensão bam com menor erro para cada amostra. Assim, o novo arquivo bam gerado foi utilizado para a busca por variações alélicas no genoma, sendo em ambos os casos, o módulo *HaplotypeCaller* utilizado para o GATK.

Estas análises foram conduzidas usando um workflow de bioinformática para análise de dados de resequenciamento NGS (LIU et al., 2014), desenvolvido no SoyKB (JOSHI et al., 2012, 2014) para busca de SNPs e Indels e foi conduzido usando XSEDE (NATIONAL SCIENCE FOUNDATION, 2014) como infraestrutura computacional, iPlant (GOFF et al., 2011) como infraestrutura de nuvem e armazenamento dos dados e sistema de workflow Pegasus (DEELMAN et al., 2005) para controlar e coordenar a gestão de dados e tarefas computacionais.

4.4. Análise de associação ampla genômicas

Para análise de associação ampla do genoma, foi utilizado o programa *GAPIT* (LIPKA et al., 2012). Um modelo linear misto foi selecionado para esta análise com um algoritmo de matriz EMMA. Além disto, foi utilizado o método de imputação de alelos maiores para dados faltantes, uma PCA total no valor de 3 e *p-value* mínimo de 7,19-e10^{-4}.

4.5. Detecção de desequilíbrio de ligação

Para medir os níveis de desequilíbrio de ligação nos acessos resistentes, foi calculado o coeficiente de correlação (r^2) dos alelos utilizando o programa *Haploview* (BARRETT et al., 2005). Os parâmetros utilizados pelo programa foram: −maxdistance 1000 −dprime –memory 2000 −minMAF 0.1 −hwcutoff 0.001. Para identificar blocos em desequilíbrio de ligação, foi incluído o parâmetro '−blockoutput GAB' no programa.

4.6. Anotação gênica, classificação funcional e predição de efeitos para importantes regiões genômicas.

Foi utilizado o programa snpEff (CINGOLANI et al., 2012) para auxiliar com a classificação funcional dos genes onde variações alélicas foram detectadas. Uma análise de enriquecimento de possíveis genes modificados devido a ação de um SNP não sinônimo foi realizada através dos websites agriGO (CHINA AGRICULTURAL UNIVERSITY, 2014) e SoyKB (JOSHI et al., 2012, 2014).

4.7. Estrutura populacional e análise de diversidade

Dados perdidos, deleções e SNPs heterozigotos foram removidos do conjunto de dados. Uma árvore filogenética do tipo neighbor-joining (NJ) foi construída através do programa MEGA5 (TAMURA et al., 2011) pelo modulo de p-distance. Um total de 4.938.168 SNPs foram utilizados para gerar uma estrutura populacional através do programa FastStructure (RAJ; STEPHENS; PRITCHARD, 2014).

Para as análises de diversidade, foram estimados a diversidade nucleotídica ($\theta\pi$) dentro da população. Para estimar $\theta\pi$, foram usados diferentes tamanhos de janelas (10 kb, 100 kb e 500 kb) sem sobreposição entre janelas adjacentes. Além disto, foi medido o coeficiente de diferenciação populacional (F_{ST}), através do conjunto de ferramentas vcftools (DANECEK et al., 2011).

4.8. Detecção de genes candidatos sob influência de seleção artificial

De acordo com os resultados estatísticos obtidos pela análise de diversidade, foi possível identificar genes sob influência de efeitos de seleção nos acessos nacionais. Regiões sob seleção positiva tendem a ter baixos valores de diversidade nucleotídicas entre acessos novos e antigos. O critério adotado para regiões sob seleção positiva foram; $F_{ST} \geq 0,45$ na distribuição populacional total e valores mais elevados de $\theta\pi$ nas cultivares mais antigas. Regiões com baixa diversidade foram determinadas através do critério de $F_{ST} \geq 0,02$. Finalmente, foi utilizado os websites Soybase (GRANT et al., 2009) , agriGo (CHINA AGRICULTURAL UNIVERSITY, 2014) e SoyKB (JOSHI et al., 2012, 2014) para realização de uma análise de enriquecimento dos genes sob influência de seleção positiva.

4.9. Identificação de variações no número de cópias (CNVs)

Para detecção de CNVs no genoma da soja, foi utilizado o programa *Copy Number estimation by a Mixture Of Poissons* (cn.MOPS), versão 1.10.0 (KLAMBAUER et al., 2012). Além disto, SoyKB foi utilizado para busca de genes inseridos nas possíveis regiões onde CNVs foram detectados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ARANZANA, M. J. et al. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. **PLoS genetics**, v. 1, n. 5, p. e60, nov. 2005.

BARRETT, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics (Oxford, England)**, v. 21, n. 2, p. 263–5, 15 jan. 2005.

BENKO-ISEPPON, A. M.; NEPOMUCENO, A. L.; ABDELNOOR, R. V. GENOSOJA – The Brazilian Soybean Genome Consortium : High throughput omics and beyond. **Genetics and Molecular Biology**, v. 35, n. 1, p. i–iv, 2012.

BORÉM, A.; MIRANDA, G. V. **Melhoramento de Plantas**. Viçosa, MG: UFV, 2005.

CAO, J. et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. **Nature Genetics**, v. 43, n. 10, p. 956–963, 2011.

CARVALHO, M. C. DA C. G. DE; SILVA, D. C. G. DA. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas Next generation DNA sequencing and its applications in plant genomics. **Ciência Rural**, v. 40, n. 3, p. 735–744, 2010.

CHINA AGRICULTURAL UNIVERSITY. **ANALYSIS TOOLKIT FOR THE AGRICULTURAL COMMUNITY (agriGO)**. Disponível em: <http://bioinfo.cau.edu.cn/agriGO/analysis.php>. Acesso em: 14 nov. 2014.

CHUNG, W.-H. et al. Population Structure and Domestication Revealed by High-Depth Resequencing of Korean Cultivated and Wild Soybean Genomes †. **DNA Research**, v. 21, n. 4, p. 153–167, 2014.

CINGOLANI, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3. **Landes Bioscience**, v. 6, n. 2, p. 80–92, 2012.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Séries Históricas de Área Plantada, Produtividade e Produção, Relativas às Safras 1976/77 a 2014/15 de Grãos, 2001 a 2014 de Café, 2005/06 a 2014/15 de Cana-de-Açúcar.** Disponível em: <http://www.conab.gov.br/conteudos.php?a=1252&>. Acesso em: 14 nov. 2014.

COOK, D. E. et al. Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean. **Science**, v. 338, n. 6111, p. 1206–1209, 6 maio 2012.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics (Oxford, England)**, v. 27, n. 15, p. 2156–8, 1 ago. 2011.

DEBOLT, S. Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. **Genome Biology and Evolution**, v. 2, p. 441–453, 2010.

DEELMAN, E. et al. Pegasus : A framework for mapping complex scientific workflows onto distributed systems. **Scientific Programming**, v. 13, n. January, p. 219–237, 2005.

EMBRAPA SOJA. **EMBRAPA SOJA. História: Histórico no Brasil.** Disponível em: <https://www.embrapa.br/en/soja/cultivos/soja1/historia>. Acesso em: 14 nov. 2014.

FERREIRA, M. E.; GRATTAPAGLIA, D. Genética de Associação de Plantas. In: BORÉM, A.; CAIXETA, E. T. (Eds.). . **Marcadores Moleculares**. Viçosa, MG: [s.n.]. p. 273–306.

GOFF, S. A et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. **Frontiers in plant science**, v. 2, n. July, p. 34, jan. 2011.

GRANT, D. et al. SoyBase, the USDA-ARS soybean genetics and genomics database. **Nucleic Acids Research**, v. 38, p. 843–846, 2009.

GU, W.; ZHANG, F.; LUPSKI, J. R. Mechanisms for human genomic rearrangements. **PathoGenetics**, v. 1, p. 4, 2008.

HAO, D. et al. Genome-wide association analysis detecting significant single nucleotide polymorphisms for chlorophyll and chlorophyll fluorescence parameters in soybean (Glycine max) landraces. **Euphytica**, v. 186, p. 919–931, 2012.

HOAGLAND, D. R.; ARNON, D. I. **The water-culture method for growing plants without soil**. [s.l: s.n.]. v. 347p. 1–32

HUANG, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature genetics**, v. 42, n. 11, p. 961–7, nov. 2010.

HWANG, E.-Y. et al. A genome-wide association study of seed protein and oil content in soybean. **BMC genomics**, v. 15, p. 1, 2014.

IAFRATE, A J. et al. Detection of large-scale variation in the human genome. **Nature genetics**, v. 36, n. 9, p. 949–951, 2004.

IQUIRA, E.; HUMIRA, S.; FRANÇOIS, B. Association mapping of QTLs for sclerotinia stem rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach. **BMC plant biology**, p. 1–12, 2015.

JOSHI, T. et al. Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. **BMC genomics**, v. 13 Suppl 1, n. Suppl 1, p. S15, jan. 2012.

JOSHI, T. et al. Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. **Nucleic acids research**, v. 42, n. Database issue, p. D1245–52, jan. 2014.

KIM, M. Y. et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb . and Zucc .) genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 51, p. 22032–22037, 2010.

KLAMBAUER, G. et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. **Nucleic acids research**, v. 40, n. 9, p. e69, maio 2012.

KLEINJAN, D. A; VAN HEYNINGEN, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. **American journal of human genetics**, v. 76, p. 8–32, 2005.

KUMP, K. L. et al. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. **Nature genetics**, v. 43, n. 2, p. 163–8, fev. 2011.

LAI, J. et al. Genome-wide patterns of genetic variation among elite maize inbred lines. **Nature genetics**, v. 42, n. 11, p. 1027–1030, 2010.

LAM, H.-M. et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. **Nature genetics**, v. 42, n. 12, p. 1053–9, dez. 2010.

LEE, G.-A. et al. Archaeological soybean (Glycine max) in East Asia: does size matter? **PloS one**, v. 6, n. 11, p. e26720, jan. 2011.

LEE, J. A.; CARVALHO, C. M. B.; LUPSKI, J. R. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. **Cell**, v. 131, p. 1235–1247, 2007.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics (Oxford, England)**, v. 25, n. 14, p. 1754–60, 15 jul. 2009.

LI, Y. et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. **BMC genomics**, v. 14, n. 1, p. 579, jan. 2013.

LIEBER, M. R. The mechanism of human nonhomologous DNA End joining. **Journal of Biological Chemistry**, v. 283, n. 1, p. 1–5, 2008.

LIPKA, A. E. et al. GAPIT: Genome association and prediction integrated tool. **Bioinformatics**, v. 28, n. 18, p. 2397–2399, 2012.

LIU, Y. et al. Large Scale NGS resequencing data analysis workflow for soybean germplasm using iPlant, XSEDE and SoyKB framework. **Bioinformatics (Oxford, England)**, v. in press, 2014.

LU, P. et al. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. **Genome Research**, v. 22, p. 508–518, 2012.

MAMIDI, S. et al. Genome-Wide Association Analysis Identifies Candidate Genes Associated with Iron Deficiency Chlorosis in Soybean. **The Plant Genome Journal**, v. 4, n. 3, p. 154, 2011.

MAMIDI, S. et al. Genome-Wide Association Studies Identifies Seven Major Regions Responsible for Iron Deficiency Chlorosis in Soybean (Glycine max). **PloS one**, v. 9, n. 9, 2014.

MASSMAN, J. et al. Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. **Molecular Breeding**, v. 27, n. 4, p. 439–454, 29 abr. 2010.

MCHALE, L. K. et al. Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. **Plant Physiology**, v. 159, n. August, p. 1295–1308, 2012.

MCKENNA, A. et al. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. p. 1297–1303, 2010.

MORRIS, G. P. et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 2, p. 453–8, 8 jan. 2013.

MUÑOZ-AMATRIAÍN, M. et al. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. **Genome biology**, v. 14, n. 6, p. R58, 2013.

NATIONAL SCIENCE FOUNDATION. **The Extreme Science and Engineering Discovery Environment (XSEDE)**. Disponível em: <https://www.xsede.org/home>. Acesso em: 14 nov. 2014.

QIN, C. et al. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 14, p. 5135–40, 8 abr. 2014.

QUTOB, D. et al. Copy number variation and transcriptional polymorphisms of Phytophthora sojae RXLR effector genes Avr1a and Avr3a. **PLoS ONE**, v. 4, n. 4, 2009.

RAJ, A.; STEPHENS, M.; PRITCHARD, J. K. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. **Genetics**, v. 197, n. June, p. 573–589, 2014.

ROVELET-LECRUX, A. et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. **Nature genetics**, v. 38, n. 1, p. 24–6, jan. 2006.

SCHMUTZ, J. et al. Genome sequence of the palaeopolyploid soybean. **Nature**, v. 463, n. 7278, p. 178–83, 14 jan. 2010.

SEBAT, J. et al. Large-scale copy number polymorphism in the human genome. **Science (New York, N.Y.)**, v. 305, n. 2004, p. 525–528, 2004.

SEBAT, J. et al. Strong Association of De Novo Copy Number Mutations with Autism. **Science**, v. 316, n. 5823, p. 445–449, 2010.

SEDIYAMA, T. et al. Genética e Melhoramento. In: **A soja no Brasil central.** 3.ed. ed. Campinas, SP: FUNDACAO CARGILL, 1986. p. 22–74.

SIMON-SANCHEZ, J. et al. Genomewide SNP Assay Reveals Mutations Underlying Parkinson Disease. **Human Mutation**, v. 29, n. 2, p. 315–322, 2008.

SPRINGER, N. M. et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. **PLoS Genetics**, v. 5, n. 11, 2009.

TAMURA, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. **Molecular biology and evolution**, v. 28, n. 10, p. 2731–9, out. 2011.

TUZUN, E. et al. Fine-scale structural variation of the human genome. **Nature genetics**, v. 37, n. 7, p. 727–732, 2005.

WETERINGS, E.; VAN GENT, D. C. The mechanism of non-homologous end-joining: A synopsis of synapsis. **DNA Repair**, v. 3, p. 1425–1435, 2004.

YU, P. et al. Detection of copy number variations in rice using array-based comparative genomic hybridization. **BMC genomics**, v. 12, n. 1, p. 372, 2011.

ZHENG, L.-Y. et al. Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). **Genome Biology**, v. 12, n. 11, p. R114, 2011.

ZHOU, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. **Nature biotechnology**, n. April 2014, 2 fev. 2015.

ZMIEŃKO, A. et al. Copy number polymorphism in plant genomes. **Theoretical and Applied Genetics**, v. 127, p. 1–18, 2014.

ARTIGO I - EVALUATION OF GENETIC VARIATION BY GENOME RESEQUENCING OF 28 BRAZILIAN SOYBEAN CULTIVARS

Abstract

Background

Soybean [*Glycine max* (L.) Merrill] is one of the most important cultivated legumes in the world, and Brazil is one of the main soybean producers. Since its reference genome was sequenced, interest has grown on structural and allelic variation in cultivated and wild soybean germplasm. In order to investigate the genetic basis of Brazilian soybean germplasm, we re-sequenced soybean cultivars selected based on release time, different geographical region and maturity groups.

Results

We re-sequenced the genome of 28 Brazilian soybean cultivars, with an average cover of 14.78X genome cover. A total of 5,868,344 SNPs and 1,329,844 indels were identified throughout the 20 soybean chromosomes, with 541,760 exclusive SNPs in 28 cultivars. In addition, 693 SNPs were shared among all Brazilian cultivars inside 321 genes, including some related to DNA-dependent transcription/elongation, cellular respiration, photosynthesis, ATP synthesis coupled electron transport, generation of precursor metabolites and energy. Furthermore, a very homogeneous structure could be observed in the Brazilian genetic basis. Additionally, we observed 57 putative regions under influence of positive selection, of which chromosome 17 had the highest number among these regions. Finally, we detected 3,880 regions with copy-number variation (CNVs) that could help to explain the divergence among the accessions used in this study.

Conclusions

The large number of allelic and structural variations identified in this study can be used in marker-assisted selection programs, as the detection of unique SNPs to cultivar fingerprint. The results suggested that despite the fact that Brazilian soybeans are diversifying in the modern cultivars, the genetic basis is still very narrow, because of the large number of regions with low diversification in the genome. These results emphasize the need to introduce new alleles from other germplasm to be used on crossing with Brazilian elite cultivars and increase the genetic diversity of Brazilian germplasm.

Keywords: *Glycine max*, allelic variation, genetic diversity, positive selection, CNVs.

Background

Soybean [*Glycine max* (L.) Merrill] is considered one of the most important leguminous crops of the world, due to its importance in human food, animal feed, and oil production. Globally, Brazil is the second largest soybean producer with potential to become the largest producer in near future. In the 2013/14 growing season Brazil harvested 86.273 million of tons, from 30.100 million hectares of sowed area [1], indicating the importance of this crop to the Brazilian agribusiness, and demonstrating the importance breeding programs that aims to the increase yield, stress tolerance and crop quality.

The Brazilian soybean breeding has a very recent history, with the first cultivar (cv.) dating from the 40s. Soybean became economically important in the 70s, and since then, it has increased the importance of this leguminous crop in the agricultural market in the world. The success of soybean for the Brazilian agribusiness is the direct result of the increase of the production in traditional areas, and the advancement of new agricultural frontiers, mainly in the Savannah region, associated to availability of genetic cultivar for tropical regions [2]. Although the great progress and achievements have been made by the soybean breeding programs has reached in Brazil, some factors are still limiting the crop potential, such as diseases and unfavorable environment conditions. The Brazilian soybean germplasm has a narrow genetic base, which restricts productivity gains, besides offering risk in cases of new pathogenic variants or emerging diseases. In previous studies, Hiromoto and Vello [3] found 26 soybean ancestors with significant contribution to the Brazilian soybean genetic bases, being PI 548485 (Roanoke), PI 548445 (CNS), PI 548493 (Tokyo), and PI 548488 (S-100) the most important ancestors. Furthermore, a new study showed that these four ancestors contributed to 55.3% of the Brazilian soybean genetic base [4]. Moreover, the same study showed that Brazilian soybeans have six important ancestors shared with US soybean genetic base (CNS, S-100, Roanoke, Tokyo, PI 54610 and PI 548318), since the first Brazilian cultivars were developed based on the US germplasm.

Therefore, the development of tools that can help breeding programs to keep the demand for cultivars with high yield and adapted to different stress conditions are essential for the increasing demand of food in the world. Molecular biology

techniques have emerged as important tools for plant breeding assistance. Thus, the new high-throughput sequencing platforms arise as alternatives to genes, quantitative trait loci (QTL) and important allelic variation discovery, as well as population studies and genome-wide association analysis (GWAS) in plants [5].

In soybeans, large-scale sequencing efforts have taken place recently, such as the genome reference sequencing [6], where 1,1 gigabase (Gb) of Williams 82 cultivar was assembled allowing the identification of 46,430 genes spread through the 20 soybean chromosomes, 70% higher than as observed in *Arabidopsis thaliana*. Furthermore, the same study showed that about 75% of the genes present in the soybean genome have multiple copies.

Wild soybean, *Glycine soja*, has also being studied at the genome level. Kim et al. [7] sequenced 915.5 megabase (Mb) of a wild soybean accession and found 2.5 megabase of substituted sequences, 406 kilobase (Kb) of indels, 32.4 megabase of deletions and 8.3 megabase of new sequences when compared with the *Glycine max* reference genome, cv. Williams 82.

Amidst the large amount of information generated by genome-wide sequencing, resequencing strategies arise as an important tool for allelic variation studies. A large number of resequencing studies in soybean can be reported. For instance, Lam et al. [5] identified a high diversity in wild soybean accessions through resequencing of 31 wild and commercial soybean cultivars. Also, the same study identified 205,614 single nucleotide polymorphisms (SNPs) that could be used in genome-wide association studies (GWAS) and QTL mapping. Chung et al. [8] cataloged genomic variation in commercial and wild soybean accessions from Korea and identified 3.87 millions of high quality SNPs. Li et al. [9] re-sequenced 25 new Chinese soybean accessions and 30 soybean accessions found in public database and identified 5,102,244 SNPs and 707,969 indels, of which 25.5% had not been reported. In other species, whole-genome resequencing has been also widely used, for instance, *Arabidopsis* [10], corn [11], rice [12], cucumber [13] and sorghum [14].

A large amount of sequence information continuously deposited in public databases shows the importance of such studies to better understand the genetic basis of this leguminous. The advent of new high-throughput sequencing technologies for genome-wide analysis, associated with a cost reduction, have allowed massive genome sequencing of large number of lines in different crops.

Thus, resequencing strategies arise as an important tool to identify variations that can be useful for breeding programs with narrow genetic basis, like soybean.

In this study, we resequenced 28 Brazilian soybean lines released over last 50 years and belonging to different maturity groups. These sequences were used to evaluate the modification among the genomes through the history of the Brazilian soybean breeding programs. Furthermore, we identified genomic regions associated with important variation, such as deletions, substitutions and duplication, which could be helpful for explaining the divergence/similarity among different cultivars.

Results and Discussion

Sequencing and variation calling

In this study, 28 Brazilian soybean accessions were re-sequenced. The Brazilian lines were chosen based on the distribution along a 50-year span of the soybean breeding program history in Brazil, with cultivars developed from the 60's until the present decade. Furthermore, lines from different maturity groups, adapted to different regions of Brazil, were also selected, representing the highest diversity among these cultivars (**Table 1**).

The resequencing effort of Brazilian cultivars generated around 5.5 billion paired-end reads with 100 bp read length and an average of coverage of 14.78x the soybean genome. The percentage of mapped reads on the soybean genome for each accession was 94.31%, which showed that the resequencing was able to cover most part of the genome (**Supplementary Table 1**). A total of 5,868,344 SNPs were identified in Brazilian lines compared to the reference genome, which is higher than previously studies [5, 8]. It was expected due to the higher coverage used in our study compared to the others. The SNPs were distributed over all the chromosomes; with chromosomes 15 and 18 having the largest number of SNPs (**Figure 1a**) and the highest ratio of SNPs per chromosome length (**Supplementary Table 2**). As expected, most of the SNPs/Indels were homozygous. However, 7.17% of them were heterozygous, being Embrapa 48 the cultivar with the highest number of heterozygous SNPs (**Supplementary Figure 1a**). When compared to the reference genome, most of the nucleotide change was classified as transition, with a transition/transversion ratio (ts/tv ratio) of 1.83 (**Figure 1b**). A total of 2,684,448 SNPs were detected in intergenic regions. In coding regions, we found a total of 218,671 SNPs in exons, 287,414 SNPs in introns and 112,790 in UTR regions

(**Figure 1c**). The non-synonymous-to-synonymous ratio observed between the Brazilian accessions was 1.55. The ratio observed in this study was lower than those observed in other soybean studies [5]; however, higher than those observed in other plants, such as sorghum [14] and rice [12]. The genomes of cv. Santa Rosa and Doko had the highest number of SNPs in coding regions, while in contrast cv. Anta 82 and VMAX RR have the lowest number (**Supplementary Table 3**). Furthermore, we identified exclusive SNPs for each of the lines used in our study. BRS 284, BRS/GO Chapadões, Doko and Santa Rosa had the highest number of exclusive SNPs, in contrast to cv. BRS Valiosa RR, BRSMG 850G RR and FT Cristalina, with the lowest number (**Table 2**). These findings can be very useful in breeding programs for marker-assisted selection (MAS) and cultivar fingerprinting.

A total of 1,329,844 indels were detected among Brazilian soybean accessions. It was lower than the proportion observed in other species [12, 14]. The distributions of indels on chromosomes, as well as the proportion of homozygous/heterozygous indels for each cultivar, were similar to those observed for SNPs (**Figures 1a and Supplementary Figure 1b**). About 463,106 of these indels were on the intergenic regions, while 79,721 indels (5.99%) were found in intronic regions, 40,105 indels (3.02%) were inside UTR regions and a total of 25,861 indels were located in exons. Similar to the SNPs analysis, Doko and Santa Rosa had the highest number of InDels, while in contrast BRS 284 and BRS/GO 8360 had the lowest number. A summary of these variations can be showed in **Figure 1d**.

Possible influence of allelic variation on the Brazilian germplasm

The SNPs identified in the Brazilian germplasm led to a large number of codon modifications detected in important gene regions. We detected 27,582 genes with the presence of allelic variation in Brazilian lines (**Supplementary Table 4**). The cultivar with the most number of modified genes due to the presence of SNPs was cv. Emgopa 301. This cultivar had the largest number of non-synonymous modifications in coding sequence regions and start codons. In contrast, we observed a small number of modified genes in cv. BR 16.

We found 17,581 SNPs that share the same allele in all Brazilian lines and are divergent to the reference genome and some of the 19 U.S. soybean germplasm. It was evident that these SNPs are common among all Brazilian germplasm used on this study but not in all U.S. cultivars. A total of 609 SNPs were non-synonymous

modifications identified in coding regions of 303 genes. According to an enrichment analysis from SoyBase [15], 34 genes were associated with generation of precursor metabolites and energy related to DNA-dependent transcription/elongation and photosynthesis biological processes. Some of them were also related to cell respiration and ATP synthesis coupled electron transport (**Supplementary Table 5**).

In addition, we found seven genes with putative modifications that cause a loss of function in start codons, shared between all Brazilian lines. These genes are related to protein biding (*Glyma.07g153200*), ATP synthesis coupled electron transport and NADH dehydrogenase (ubiquinone) activity (*Glyma.15g246000*) and three putative pseudogenes on chromosome 16 (15,19-16,88 Mb): *Glyma.16g017300* (a serine/threonine protein kinase), *Glyma.16g019100* (proprotein convertase subtilisin/kexin) and *Glyma.16g019200* (S1/P1 Nuclease related with DNA catabolic processes). In stop codons, we identified six allelic variations, but only two genes had annotation. The *Glyma.07g156200* has an AP2 domain, related to the transcription regulation and *Glyma.18g132800* is associated with the ATP binding, being a chloroplast cell component.

Moreover, we detected five genes with modifications on the splice site regions, leading to alternative splicing, including a gene with PPR repeat domain (*Glyma.18g056000*) that could be related with plant resistance mechanism, a NADH-Ubiquinone/plastoquinone (*Glyma.10g068800*), a regulation of DNA replication (*Glyma.16g005600*) and mRNA splicing process (*Glyma.16g077700*). In contrast, there is no annotation for gene *Glyma.17g186300*.

Finally, for 14,560 SNPs the same allele was found in all Brazilian cultivars and all the U.S. accessions, comparing to cv. Williams 82. One of the U.S. soybean, US-08, was originated by a cross among cv. Williams 82 and Kin du, which suggest that the presence of these SNPs were sequencing errors on the reference genome or SNPs that are unique to cv. Williams 82. A total of 760 SNPs were observed inside coding regions of 476 genes, with 24 also associated with generation of precursor metabolites and energy related to DNA-dependent transcription/elongation and photosynthesis biological processes. Moreover, we observed 36 SNPs in splice site, start and stop codon regions of 33 genes. This result may have a large importance in future soybean studies, due the presence of possible nucleotide errors in important genes of the genome reference. In order to confirm the existence of these variations, a validation procedure must be carried out for these loci.

Several gene modifications were found in Brazilian accessions compared to the reference genome and other U.S. lines. Once confirmed, these differences could be an insight about the adaptation to the tropics in Brazil and some of the differences that led to loss of function may not have a key role to the plant survival. However, more detailed studies are needed to verify the importance of these modified genes, especially in photosynthesis and generation of precursor metabolites and energy processes. Whether these genes were involved in tropical adaptation of soybean, it still needs to be clarified.

Low divergence in Brazilian soybean genetic base

The Brazilian soybean genetic base is very narrow due to the very recent breeding program history and the existence of a small number of ancestors, mainly derived from the United States (US) soybean germplasm. To observe the population structure of the Brazilian soybean germplasm, we constructed a neighbor-joining (NJ) tree with all the Brazilian soybean cultivars (**Figure 2a**). The phylogenetic tree showed that the accessions were grouped according to their genealogy. It was possible to confirm it through the analysis of some clusters.

Cultivar BRS Sambaíba has a strong influence of Santa Rosa on its pedigree, which explains these two cultivars were clustered together. Moreover, cv. Paraná and IAS 5 have a great influence of Hill and D52-810, which explains the clustering between these two genotypes. Hill was present at least in 19 Brazilian cultivars background and should have an importance in the clustering. Similar situations occurred between cv. FT Cristalina and cv. BRSMT Uirapuru, because the latter has the former background in its pedigree. Cultivars BRS 284 and BRSGO 8360 share the same common ancestors (Mycosoy-45 x Suprema), which also explains why these two accessions were clustered together in the phylogenetic tree. Moreover, BRS 361 was clustered close to BRS 284 due the presence of cv. Suprema in its background. Emgopa 301 and BRS 284 shared Hardee and Improved Pelican as common ancestors, which explain the clustering of these cultivars. The cultivars BR/MG 46 (Conquista) and BRS Valiosa RR also have a high genealogical relationship, since BRS Valiosa RR was originated from five backcrosses of cv. Conquista. Another important cultivar, BRSMG 850G RR, was closely clustered to cultivars BRS Valiosa RR and Conquista, which can be explained by their similar backgrounds. BRS 232 and FT Abyara had Hood and Hill in their background, which

should have a great influence in their position in the phylogenetic tree. Furthermore, cultivars IAC 8 and Conquista shared Bragg-derived genetic background as a common ancestor. Priolli et al [16] clustered 168 Brazilian soybeans in a UPGMA tree based on SSR polymorphism and identified a similar clustering. Moreover, the same study clustered by Roger's Wright distance, put BR 16 close to FT Cristalina, which could explain the closeness of these two cultivars.

In addition, some resistant cultivars against soybean cyst nematode were clustered closely. It was explained due the influence of common resistant ancestral in their background. This fact can be more clarified with the presence of BRS 360RR at the same cluster, since two U.S. cultivars commonly used as source of resistance against soybean cyst nematode are present in its background, even BRS 360RR being a susceptible cultivar.

Using FastStructure [17], we check the homogeneity of the Brazilian soybean genetic basis. The *K* value was set from a range between 1 to 10 and the best model components used to explain structure in this data was the model *K* = 3 (**Figure 2b)**. Based on the phylogenetic tree, most of the samples were clustered according to their background influence. Moreover, some evidence of admixture can be seen in BR 16, BRS 232, BRS 361, BRS/GO 8660, BRS/GO Chapadões, Doko, EMGOPA 301, FT Abyara and Santa Rosa. The results suggest that the Brazilian soybean genetic basis is still very homogeneous, with possible introgression in few cultivars. Priolli et al [18] used 435 cultivars and 27 SSR markers to cluster Brazilian soybeans in two groups (k=2). The main reason of the clustering difference is involved in the number of cultivars and markers used in both studies, since we used a lower number of cultivars but a higher number of markers.

The breeding programs develop cultivars with the best performance under influence of various environment and field conditions. Against this background, the development of cultivars tends to modify some genes/QTLs through the time by the increase/removal of important alleles in the genome. Thus, the identification of regions with higher diversification, as well as regions with lower modifications during the time, become extremely important for breeding programs to improve the soybean adaptation.

In order to identify genomic regions with higher levels of diversification between old and new accessions, a calculation of the divergence index (F_{ST}), between the Brazilian accessions was performed. Regions with higher F_{ST} values

could be related to artificial selection events, as well as regions with lower values of F_{ST} could indicate the existence of little genetic differentiation between the accessions.

In this study, we identified 998 10 kb regions with F_{ST} values higher than 0.45 distributed in most of the soybean chromosomes. The chromosome 16 had the higher number of sub-region with higher values of F_{ST}. Two chromosomes, 9 and 13, did not present sub-regions with higher values of F_{ST}, which may mean these chromosomes do not have strong influence of artificial selection during the development of new cultivars.

In contrast, we detected 2,097 sub-regions with F_{ST} values lower than 0.02, which showed the higher number of genomic regions with lower diversification between latest and oldest cultivars. Chromosome 6 had the highest number of these sub-regions with lower diversification, while chromosome 16 had the lowest number. Lam et al [5] identified 369 sub-regions with high F_{ST} and 101 sub-regions with low F_{ST} in a comparison between wild and commercial soybeans. The proportion of higher/lower F_{ST} detected in this study was higher compared to our results. It can be explained since the previous study compared *Glycine soja* and *Glycine max* accessions, while in our study, only commercial accessions of the same geographic region was used. A large number of sub-regions with lower diversification related to the population structure analysis results obtained in this study demonstrated that the genetic basis of the Brazilian soybean germplasm remains narrow. These observations are in agreement with those studies, in which 444 Brazilian soybean lines were investigated and showed the same pattern [4]. In that study, it was identified a cumulative relative genetic contributions of 57.6% from only four main ancestors in all Brazilian germplasm, even with an increase of the number of ancestors in the genetic basis over the time.

High similarity between Brazilian and U.S. soybean genetic base

The history of the soybean breeding program in Brazil is very recent and associated with the U.S. breeding programs, as the first soybean cultivars used in Brazil were introduced from the United States. Therefore, it is expected that the Brazilian and U.S. genetic basis to be very close. In order to verify genomic regions that were modified over the time, comparisons of the population differentiation coefficient (F_{ST}) were made between Brazilian and U.S. soybean cultivars.

A total of 261 sub-regions with high values of F_{ST} were identified between Brazilian and U.S. accessions. There are no sub-regions on chromosome 4 and 8 with higher values of F_{ST}. Chromosome 6 with 78 and chromosome 10 with 74 had the highest number of sub-regions with higher values of F_{ST}. One region, on chromosome 6, between 0.63-0.69 Mb had the highest number of SNPs with great values of F_{ST}. There are seven genes with important modifications in coding sequence in this region: *Glyma.06G00820* (endo-1,4-beta-glucanase, related with carbohydrate metabolic process), *Glyma.06G008300* (RNA polymerase I transcription factor UAF), *Glyma.06G008400* (iron/ascorbate family oxidoreductases associated with oxidation-reduction process), *Glyma.06G008500* (E3 ubiquitin ligase interacting with arginine methyltransferase), *Glyma.06G008800* (a transcription factor GATA related with regulation of transcription), *Glyma.06G008600* and *Glyma.06G008700*. QTLs for oil/palmitic acid content [12], carbon isotope discrimination [19] and reaction to *Heterodera glycines* damage [20] have been previously described inside this region, that could be an important source of variation between both genetic bases, especially because it is related to important soybean traits and have higher diversification between both germplasm.

In contrast, we observed 8,167 sub-regions with F_{ST} values lower than 0.02. Chromosome 19 had the highest number of detected sub-regions, while only 151 sub-regions were found on chromosome 16 had the lowest number, with 151 sub-regions. The large numbers of sub-regions with lower diversification in Brazilian and U.S. soybeans support the idea of both genetic bases are still very close.

Following the analysis, a comparison between the recent lines from both genetic bases were made as a way to verify the existence of regions more influenced over time by the process of diversification. We found 900 sub-regions with F_{ST} values higher than 0.45 and 1,958 sub-regions with F_{ST} values lower than 0.02. The results obtained in this analysis suggest that although the genetic bases of both accessions are close, a genetic diversification occurred between the latest cultivars. The chromosomes 6 and 18, with 208 each one, were identified with the highest number of sub-regions with F_{ST} values higher than 0.45.

Regions under positive or balancing selection processes in Brazilian germplasm

A large number of sub-regions with the highest values for F_{ST} associated with lower values of nucleotide diversity (θπ) for the latest cultivars compared to the oldest cultivars were identified on chromosomes 7, 15, 17, and 18 (**Table 3**). We identified 32 sub-regions of 10 kb size inside two intervals on chromosome 17, being 8 sub-regions inside the interval 3.01-3.06 Mb with 100 SNPs and 39 sub-regions between 5.56-5.92 Mb with 1,150 SNPs (**Figure 3**). Most of the SNPs identified in both intervals differentiated Doko, IAC 8, IAS 5 and Paraná from the other cultivars. These intervals were previously described in other studies by the presence of a large number of QTLs, such as seed size [21–24], seed genistein/palmitic acid content [25, 26], plant/root weight, phosphorus content [27], canopy wilt [28], soybean cyst nematode [29] and white mold [30]. Moreover, we identified 30 genes with allelic variation in important regions with variable functions, such as late embryogenesis abundant [plants] lea-related (*Glyma.17G040800*), a ribosomal protein L35 (*Glyma.17G07120*), x-box transcription factor-related with cellulose biosynthetic process (*Glyma.17G072200*), a gene with a WRKY DNA-binding domain related with the regulation of transcription (*Glyma.17G074000*), a RNA polymerase II CTD phosphatase (*Glyma.17G074700*), two ABC transporter (*Glyma.17G041300* and *Glyma.17G041200*), three Zinc-finger double-stranded RNA-binding (*Glyma.17G072800*, *Glyma.17G072900* and *Glyma.17G073000*), a gene with a PPR repeat domain (*Glyma.17G072100*) and a gene related with cytochrome-c oxidase activity (*Glyma.17G075100*).

We also identified additional sub-regions on chromosomes 7, 15, and 18 with higher values of F_{ST}. Six sub-regions, located by the end of chromosome 7, were detected and all these sub-regions carried SNPs that differed cultivars IAC 8, Santa Rosa, and Doko from the other cultivars. Tajuddin et al. [31] described two QTLs for seed-oil content inside these sub-regions. In the present study, we identified four genes between 40.10-40.17 Mb: *Glyma.07G223900* (DNA helicase PIF1/RRM3, associated with telomere maintance), *Glyma.07G224100* (a gene with a B3 DNA binding domain), *Glyma.07G224400* (NusB family associated with the regulation of transcription) and *Glyma.07G224600* (a glucosidase 2 subunit beta). On the other hand, three other sub-regions detected on chromosome 15 (2.95-2.97 Mb, with 51 SNPs) and 18 (2.19-2.20 Mb, with 107 SNPs) were in the beginning of these

chromosomes. Only in chromosome 18 we identified a modified gene due to the existence of a SNP: *Glyma.18G029000*, an amino acid transporter. However, several studies reported the presence of QTLs controlling important traits in these sub-regions. On chromosome 15, several QTLs responsible for seed volume/length [21], isoflavone content [32], oleic/linoleic acid content [33] and protein/oil content [31, 34] were identified. For chromosome 18, most of the QTLs identified were related to soybean cyst nematode resistance [35–42] and protein content [43]. The SNPs found in this study on chromosome 15 differentiated IAC 8, Paraná, and Doko from the latest cultivars. However, we identified a similar pattern in cv. Embrapa 48 compared to the oldest cultivars. This could be explained by the presence of Paraná in its pedigree. Furthermore, in chromosome 18, the SNPs identified in this study differentiated IAS 5, Paraná, and Doko from the latest cultivars.

The higher F_{ST} values associated with higher $\theta\pi$ values in oldest cultivars compared to the latest cultivars confirmed the presence of sub-regions under positive selection processes. Thus, the Brazilian accessions had meaningful modifications in these 41 sub-regions over the time. The presence of important traits inside these sub-regions associated with the large difference in Brazilian production over the time and high F_{ST} values reinforces the existence of sub-regions under influence of positive selection.

We also identified a large number of regions with F_{ST} under 0.02. This result suggests the presence of regions with lower diversification, which could mean the presence of balancing selection. Part of these regions under balancing selection could have important genes/QTLs responsible for the survival of the plant. This finding, associated with the detection of a large number of regions with higher FST values, could be an important target for breeding programs, to maintain these regions under positive selection. Moreover, the identification of the regions under balancing selection not related to essential processes for the plant could be another important target to insert new alleles that could improve important traits in Brazilian cultivars.

Copy-Number Variation regions could explain the divergence among accessions

CNVs refer to structural modifications that produce changes in the copy number in a specific region of the genome. Such modifications may vary in size, and recently some studies show their wide importance due to the fact they are linked to

many types of traits and some diseases, such as Alzheimer [44], autism [45] and Parkinson [46] in humans. In soybean, it is also observed that a significant number of CNVs are associated to important traits, as the resistance against soybean cyst nematode [47]. Therefore, the identification of these CNVs on soybean genome becomes extremely important. We analyzed all the Brazilian soybean lines to find important CNVs that could be related to the divergence accumulated during the time between the oldest and latest accessions.

A total of 3,880 sub-regions containing CNVs, spread through the 20 chromosomes of the soybean genome, were detected in the Brazilian lines. The highest number of CNVs regions were identified on chromosome 14 and 17 and the lowest number on chromosome 16. Cultivar BRS 284 has the highest number of CNVs, in contrast to BRS/GO Chapadões, a cultivar with lowest number of CNVs. Moreover, cv. BRS 284 had virtually an equal proportion of CNVs in the genome, with 671 deletions and 667 duplications. Most of the accessions had more deletions than duplications, except for BRS/GO 8360 and VMAX RR, which had more duplicated regions instead of deletions. BRS 232, Doko, and Santa Rosa were cultivars with the highest number of CNV regions. This result was similar to the previously SNP analysis described in this study. Cultivar VMAX RR was the accession with the lowest number of deleted regions. Additionally, BRS 284, BRS/GO 8360, and VMAX RR were cultivars with the highest number of CNVs region with duplication, which contrast with BRS Valiosa and FT Cristalina, with the lowest number. A summary of the number of CNVs detected for the cultivars is shown in **Supplementary Figure 2**.

When comparing the oldest to the latest cultivars, the chromosome 16 presented CNVs in 12 sub-regions (**Figure 4**). More than 80% of the latest cultivars do not have these deletions, only present in the oldest cultivars Doko, EMGOPA 301, FT Abyara, IAS 5, Paraná, and Santa Rosa. There was one of these regions, ranging on 26.20-26.21 Mb, which was not found in any cultivar developed after 2000. Furthermore, this CNV was not present in more than 70% of the accessions prior to 1999. These results suggest that these 12 sub-regions identified on chromosome 16, especially the latest described, were inserted more recently in the breeding process.

Other important CNVs regions that distinguished the oldest soybean lines from the latest ones were detected on chromosomes 6, 7, 8, 9, 13, 15, and 17 (**Supplementary Figures 3**). Five meaningful deleted regions shared with more than

70% of the latest cultivars were detected on chromosome 15 between 41.37-42.68 Mb. Cultivars IAC 8, IAS 5, Paraná, Santa Rosa, Doko, and FT-Abyara have common insertions for four CNVs regions. Moreover, there are six more accessions with these insertions: BR-16, MG/BR46, BRS 232, BRS Sambaíba, BRS Valiosa RR and BRMG 850G RR. These lines shared the common ancestry with the oldest accession that we study with, which could explain the presence of these regions in these accessions. The existence of these patterns could mean a duplicated region present in oldest cultivars and deleted from the latest cultivars.

Furthermore, chromosome 7 showed relevant results. Five sub-regions between 11.60-12.44 Mb were detected with deletions present only in oldest cultivars Doko, IAS 5, Paraná, and in four of the latest cultivars: BRS 361, BRS/GO 8660, BRS/GO Chapadões and VMAX RR. Moreover, a deletion identified between 40.60-40.62 Mb was detected only in cv. Doko, Santa Rosa, and IAC 8. All accessions produced in the 1981-2000 period did not have this last CNV, which may suggest this sub-region was introgressed into the Brazilian soybean germplasm as from 1980.

Moreover, we identified important deletions in the oldest accessions and in a few recent lines on chromosomes 6, 9, and 13. Chromosome 6 had three deletions present in Doko, IAC 8, Paraná, Conquista, and three latest lines: Anta 82, BRS Valiosa RR and BRSMG 850G RR. Cultivar CD 201 presents an insertion for the same region. Thus, more than 78 % of the Brazilian accessions produced after the 70s had an introgression of these three regions in their genomes over the time. Chromosome 9 has a deletion of 8 kb present in CD 201, IAS 5, Paraná, Santa Rosa and less than 30% of the latest cultivars. Only four recent lines (Anta 82, BRS 232, BRS/GO 8360, and BRS Sambaíba) showed the same pattern observed as in the oldest cultivars. Thus, it is possible these sub-regions were introgressed in most of the latest accessions, except for those which we identify the CNVs. Finally, chromosome 13 showed important deletions in the oldest cultivars Doko, IAC 8, IAC 5, and Paraná. This fact could mean the presence of an introgression in soybeans produced after 70s.

A similar analysis was also made to compare Brazilian and U.S. genetic basis. We identified 8 sub-regions on chromosomes 3, 4, 5, 6, 9, 14 with patterns of CNVs that distinguished the Brazilian and U.S. genetic basis (**Supplementary Figure 4**). In these regions, Anta 82 was a Brazilian cultivar with the most number of

detected CNVs, being in most of the times, duplicated regions. As we saw previously in SNP diversity analysis, most of the CNVs regions were detected in chromosome 6.

On chromosome 9, a deletion located between 22.09-22.11 Mb was present in almost all of the U.S. lines, except for US-18. This deletion is completely absent in all Brazilian soybeans and it could be an important differentiated region in both genetic bases. Similar results were found on chromosomes 3, 4, 5, 6, and 14. Chromosome 3 had meaningful modifications between 16.53-16.55 Mb. We detected a deletion shared among 12 U.S. lines, BRS 284 and BRS Sambaíba. In the same region, we identified duplications of 8-10 kb in BRS/GO 8660, BRS/GO Chapadões, Doko, Embrapa 48, US-18, US-08, US-13, P98Y11 and Paraná. None modifications were found in US-06, US-07, US-09, US-17, and the remaining Brazilian soybeans.

On chromosomes 4, we observed two deletions in almost all the U.S. accessions, except to US-18. Moreover, we also identified an insertion of 8 kb present in cultivar Anta 82, a Brazilian latest accession. Similar deletion for the same accessions and an insertion for Anta 82 were also observed on chromosome 14. These results suggested the presence of duplications processes of these sub-regions during the development of this cultivar in these two genomic regions. Meanwhile, we found another deletion in almost all U.S. accessions on chromosome 5, except to the accessions US-03, US-11, US-17, and US-18. Furthermore, there was an insertion present in BRS/GO 8360, BRS/GO Chapadões, BRSMT Uirapuru, and Emgopa 301. US-11 is the oldest accession used in this study. The absence of this deletion in US-11 could be this region was present in the first U.S. soybeans and it was lost in most of the recent U.S. and Brazilian soybeans.

Finally, there were three important CNVs detected on chromosome 6. Most of the modifications observed in Brazilian soybeans were related with deletions of these regions. In contrast, inserted regions were observed in most of the U.S. soybean genome. As we observed on chromosome 3, cultivars BRS 284 and BRS 361 had inserted sequences fragments for the first two regions. Moreover, in all CNVs regions identified in this chromosome, we also detected the same deletion that was observed in most of the Brazilian cultivars for the four oldest U.S. accessions used in this study, except the last region, that we found a deletion in US-17. This result suggests these two regions were absent in both genetic basis, inserted to U.S. soybean during the 80’s and just present in Brazilian genetic basis recently.

CNVs analysis showed as an important tool to verify meaningful modifications in genomes. The detection of this modified regions have a great impact to soybean genomic studies, which can be the focus of a future study to check the importance of the gain/loss of these regions in QTL and genes.

Conclusions

This is the first study involving a historic genomic analysis of the allelic and structural variations presents in Brazilian soybean cultivars. Our results confirmed the hypothesis that the Brazilian genetic basis stills narrow and closely to U.S. genetic basis. However, it was possible to detect the presence of SNPs and CNVs that distinguished the cultivars used in this study, as well as the Brazilian germplasm from the U.S. germplasm.

Based on the comparison among the Brazilian cultivars, it was possible to confirm that a large number of the allelic modifications were observed in genes associated to generation of precursor metabolites and energy related to DNA-dependent transcription/elongation and photosynthesis. Such modifications possibly are related to important roles for adaptation of soybean in Brazil. Furthermore, the existence of a large amount of CNVs regions that allow the differentiation among the Brazilian germplasm also appears as a potential target for studies of important agronomic traits. Therefore, the further analysis of these CNV regions should be treated as a top priority in future analyzes.

The sub-regions with low diversification identified in Brazilian soybean cultivars may be regions that have not been used in breeding programs until now. These sub-regions may represent targets for incorporation of new agronomically relevant alleles. In addition, measures to increase the diversity of the Brazilian genetic bases should be considered; for example, using genotypes from different geographical regions, such as Asian germplasm, or by selecting parental genotypes more divergent for specific genome regions.

Finally, a large number of exclusive SNPs were identified for each Brazilian soybean cultivars. These results may become an important breeding tool for cultivar fingerprinting. However, a validation process will be necessary to confirm our results.

Conflict of Interests

The authors declare that they have no conflict of interests

Acknowledgements

We greatly appreciate financial support from the Coordination for the Improvement of Higher Level or Education program (CAPES). We thank the plant biotechnology and bioinformatics laboratory members at Embrapa Soja in Brazil for supporting this study. Furthermore, we thank the Molecular Genetics & Soybean Genomics Laboratory (Division of Plant Sciences) and Digital Biology Laboratory (Computer Sciences Department) at the University of Missouri in the United States for supporting the doctoral student exchange program and this research

Material and Methods

Plant accessions and sequencing

Twenty-eight Brazilian soybean cultivars were selected for this study. The cultivars were selected based on commercial released date and on the relative maturity group (RMG). The seeds were obtained from the germplasm bank of Embrapa Soja or from commercial seeds Young leaf tissue sample of each 28 Brazilian cultivars was collected at V3 growth stage. The genomic DNA for each sample was isolated with the Qiagen Mini Plant DNeasy kit (Qiagen Inc., Valencia, CA, USA), following the manufacturer's instructions. The DNA sequencing was performed at FASTERIS Company, Switzerland, on a Illumina Hiseq 2000 platform, generating 100 bp paired-end reads, with an expected coverage of 15x the soybean genome. Sequence data from 19 U.S. soybean lines, kindly provided by the Molecular *Genetics* and *Soybean* Genomics Laboratory, the University of Missouri, and were used for diversity and CNVs comparisons. The U.S. soybean lines used in diversity analysis correspond to cultivars prominent in the U.S. breeding programs, from different maturity groups and developed from 1951 until now (**Table 1**).

SNPs and Indels detection

The reads generated by the Brazilian soybean accessions resequencing were mapped to the new version of the soybean reference genome (Gmax_275_Wm82.a2.v1) through the alignment program Burrows-Wheeler Aligner (BWA) [48]. After mapping, the aligned reads were processed through Piccard tools version 1.107 to remove duplicate values and a binary file of extension bam,

representing the assembled genome of each resequenced species was generated. For SNPs/indels calling, we used the Genome Analysis Toolkit (GATK) version 3.0 [49]. This toolkit was utilized to make a local realignment in indels region and a qualitative recalibration for the purpose to generate a bam file with fewer errors for each sample. Thus, the new bam files generated were used to SNPs/indels calling of the genome. In both cases, we used the HaplotypeCaller module of the GATK.

The analysis was conducted using the bioinformatics NGS resequencing data analysis workflow [50] developed in SoyKB for SNP and Indel calling and was conducted using XSEDE [51] as the computing infrastructure, iPlant as the data and cloud infrastructure [52], and the Pegasus workflow systems [53] to control and coordinate the data management and computational tasks.

Copy-Number Variation (CNV) identification

For CNVs detection on soybean genome, we used Copy Number estimation by a Mixture Of Poissons (cn.MOPS), version 1.10.0 [54]. Furthermore, we used the SoyKB [55, 56] website to check the existence of modified genes inside the CNV regions detected.

Genetic annotation, functional classification and prediction effect for important genomic regions.

We used SnpEff program [57] to help with the functional classification of the genes where allelic variations had been detected. An enrichment analysis of these modified genes detected through SnpEff were made through the website SoyBase [15] and agriGO [58].

Population structure and diversity analysis

Missing data, deletions and heterozygous SNPs were removed from the dataset. A neighbor-joining phylogenetic tree was constructed by MEGA5 software [59] through the p-distance module. A total of 4,938,168 SNPs were used to generate a population structure by FastStructure software [17].

For the diversity analysis, we estimated the average pairwise divergence within a population ($\theta\pi$). To estimate $\theta\pi$, we used different sliding windows of different sizes (10kb, 100 kb and 500 kb) without overlap between adjacent windows.

Furthermore, we measured the population differentiation coefficient (F_{ST}), using vcftools [60].

Detection of candidate genes under influence of artificial selection

According to the statistical results obtained by the diversity analysis, we detected some candidate genes under the influence of selection effect on Brazilian accessions. Regions under positive selection tend to have low values of diversity and low allelic frequency between new and old accessions. The criterion adopted for region with positive selection was; $F_{ST} \geq 0.45$ on total population distribution and higher values of $\theta\pi$ in old cultivars. For regions with lower diversity, we adopted the criterion of $F_{ST} \geq 0.02$. Finally, we used AgriGO [58] and SoyKB website [55, 56] to make an enrichment analysis of the genes detected under positive selection influence.

References

1. **Séries Históricas de Área Plantada, Produtividade e Produção, Relativas às Safras 1976/77 a 2014/15 de Grãos, 2001 a 2014 de Café, 2005/06 a 2014/15 de Cana-de-Açúcar.** [http://www.conab.gov.br/conteudos.php?a=1252&]
2. **EMBRAPA SOJA. História: Histórico no Brasil.** [https://www.embrapa.br/en/soja/cultivos/soja1/historia]
3. Hiromoto DM, Vello NA: **The genetic base of brazilian soybean (Glycine Max) cultivars**. *Genet Brazilian J* 1986, **IX**:295–306.
4. Wysmierski PT, Vello NA: **The genetic base of Brazilian soybean cultivars : evolution over time and breeding implications**. *Genet Mol Biol* 2013, **36**:547–555.
5. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS-M, Zhang G: **Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.** *Nat Genet* 2010, **42**:1053–9.
6. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, et al.: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**:178–83.
7. Kim MY, Lee S, Van K, Kim T, Jeong S, Choi I, Kim D-S, Lee Y-S, Park D, Ma J, Kim W-Y, Kim B-C, Park S, Lee K-A, Kim DH, Kim KH, Shin JH, Jang YE, Kim K Do, Liu WX, Chaisan T, Kang YJ, Lee Y-H, Kim K-H, Moon J-K, Schmutz J, Jackson SA, Bhak J, Lee S-H: **Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb . and Zucc .) genome**. *Proc Natl Acad Sci U S A* 2010, **107**:22032–22037.

8. Chung W-H, Jeong N, Kim J, Lee WK, Lee Y un-G, Lee S-H, Yoon W, Kim J-H, Choi I-Y, Choi H-K, Moon J-K, Kim N, Jeong S-C: **Population Structure and Domestication Revealed by High-Depth Resequencing of Korean Cultivated and Wild Soybean Genomes †**. *DNA Res* 2014, **21**:153–167.
9. Li Y, Zhao S, Ma J, Li D, Yan L, Li J, Qi X, Guo X, Zhang L, He W, Chang R, Liang Q, Guo Y, Ye C, Wang X, Tao Y, Guan R, Wang J, Liu Y, Jin L, Zhang X, Liu Z, Zhang L, Chen J, Wang K, Nielsen R, Li R, Chen P, Li W, Reif JC, et al.: **Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing.** *BMC Genomics* 2013, **14**:579.
10. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D: **Sequencing of natural strains of Arabidopsis thaliana with short reads.** *Genome Res* 2008, **18**:2024–33.
11. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2007, **51**:910–8.
12. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W: **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nat Biotechnol* 2012, **30**:105–11.
13. Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, Du Y, Li Y, Lin T, Yuan J, Yang X, Chen J, Chen H, Xiong X, Huang K, Fei Z, Mao L, Tian L, Städler T, Renner SS, Kamoun S, Lucas WJ, Zhang Z, Huang S: **A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity.** *Nat Genet* 2013, **45**:1510–5.
14. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frère C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J: **Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum.** *Nat Commun* 2013, **4**:2320.
15. Grant D, Nelson RT, Cannon SB, Shoemaker RC: **SoyBase, the USDA-ARS soybean genetics and genomics database**. *Nucleic Acids Res* 2009, **38**:843–846.
16. Priolli RHG, Pinheiro JB, Zucchi MI, Bajay M, Vello NA: **Genetic Diversity among Brazilian Soybean Cultivars Based on SSR Loci and Pedigree Data**. *Brazilian Arch Biol Technol* 2010, **53**(June):519–531.
17. Raj A, Stephens M, Pritchard JK: **FastSTRUCTURE: Variational inference of population structure in large SNP data sets**. *Genetics* 2014, **197**(June):573–589.
18. Priolli HRG, Wysmierski PT, Cunha CP da, Pinheiro JB, Vello NA: **Genetic structure and a selected core set of Brazilian soybean cultivars**. *Genet Mol Biol* 2013, **36**:382–390.
19. Specht JE, Chase K, Macrander M, Graef GL, Chung J, Markwell JP, Germann M, Orf JH, Lark KG: **Soybean Response to Water : A QTL Analysis of Drought Tolerance**. *Crop Sci* 2001, **41**:493–509.
20. Wang D, Arelli PR, Shoemaker RC, Diers BW: **Loci underlying resistance to Race 3 of soybean cyst nematode in Glycine soja plant introduction 468916**. *Theor Appl Genet* 2001, **103**:561–566.

21. Salas P, Oyarzo-Llaipen JC, Wang D, Chase K, Mansur L: **Genetic mapping of seed shape in three populations of recombinant inbred lines of soybean (Glycine max L. Merr.).** *Theor Appl Genet* 2006, **113**:1459–66.
22. Gai J, Wang Y, Wu X, Chen S: **A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean**. *Front Agric China* 2007, **1**:1–7.
23. Zhang W-K, Wang Y-J, Luo G-Z, Zhang J-S, He C-Y, Wu X-L, Gai J-Y, Chen S-Y: **QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers.** *Theor Appl Genet* 2004, **108**:1131–9.
24. Mian MAR, Bailey MA, Tamulonis JP, Shipe ER, Carter Jr. TE, Parrott WA, Boerma HR: **Molecular markers associated with seed weight in two soybean populations**. *Theor Appl Genet* 1996, **93**:1011–1016.
25. Smallwood CJ: **Detection of Quantitative Trait Loci for Marker- Assisted Selection of Soybean Isoflavone Genistein**. 2012:1–93.
26. Hyten DL, Pantalone VR, Saxton AM, Schmidt ME, Sams CE: **Molecular Mapping and Identification of Soybean Fatty Acid Modifier Quantitative Trait Loci**. *J Am oil Chem Soc Am oil Chem Soc* 2004, **81**:1115–1118.
27. Liang Q, Cheng X, Mei M, Yan X, Liao H: **QTL analysis of root traits as related to phosphorus efficiency in soybean.** *Ann Bot* 2010, **106**:223–34.
28. Abdel-Haleem H, Carter TE, Purcell LC, King CA, Ries LL, Chen P, Schapaugh W, Sinclair TR, Boerma HR: **Mapping of quantitative trait loci for canopy-wilting trait in soybean (Glycine max L. Merr).** *Theor Appl Genet* 2012, **125**:837–46.
29. Yue P, Sleper DA, Arelli PR: **Mapping Resistance to Multiple Races of Heterodera glycines in Soybean PI 89772**. *Crop* 2001, **41**:1589–1595.
30. Arahana VS, Graef GL, Specht JE, Steadman JR, Eskridge KM: **Identification of QTLs for Resistance to Sclerotinia sclerotiorum in Soybean**. *Crop Sci* 2001, **41**:180–188.
31. Tajuddin T, Watanabe S, Yamanaka N, Harada K: **Analysis of Quantitative Trait Loci for Protein and Lipid Contents in Soybean Seeds Using Recombinant Inbred Lines**. *Breed Sci* 2003, **53**:133–140.
32. Gutierrez-Gonzalez JJ, Wu X, Zhang J, Lee J-D, Ellersieck M, Shannon JG, Yu O, Nguyen HT, Sleper D a: **Genetic control of soybean seed isoflavone content: importance of statistical model and epistasis in complex traits.** *Theor Appl Genet* 2009, **119**:1069–83.
33. Diers BW, Shoemaker RC: **Restriction Fragment Length Polymorphism Analysis of Soybean Fatty Acid Content1 ~ A**. *J Am oil Chem Soc Am oil Chem Soc* 1992, **69**:1242–1244.
34. Shibata M, Takayama K, Ujiie A, Yamada T, Abe J, Kitamura K: **Genetic relationship between lipid content and linolenic acid concentration in soybean seeds**. *Breed Sci* 2008, **58**:361–366.
35. Arriagada O, Mora F, Dellarossa JC, Ferreira MFS, Cervigni GDL, Schuster I: **Bayesian mapping of quantitative trait loci (QTL) controlling soybean cyst nematode resistant**. *Euphytica* 2012, **186**:907–917.
36. Vuong TD, Sleper D a, Shannon JG, Nguyen HT: **Novel quantitative trait loci for broad-based resistance to soybean cyst nematode (Heterodera glycines Ichinohe) in soybean PI 567516C.** *Theor Appl Genet* 2010, **121**:1253–66.

37. Wu X, Blake S, Sleper D a, Shannon JG, Cregan P, Nguyen HT: **QTL, additive and epistatic effects for SCN resistance in PI 437654.** *Theor Appl Genet* 2009, **118**:1093–105.
38. Ferdous SA, Watanabe S, Suzuki-Orihara C, Tanaka Y, Kamiya M, Yamanaka N, Harada K: **QTL Analysis of Resistance to Soybean Cyst Nematode Race 3 an Soybean Cultivar Toyomusume**. *Breed Sci* 2006, **56**:155–163.
39. Guo B, Sleper D a, Arelli PR, Shannon JG, Nguyen HT: **Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763.** *Theor Appl Genet* 2005, **111**:965–71.
40. Glover KD, Wang D, Arelli PR, Carlson SR, Cianzio SR, Diers BW: **Near Isogenic Lines Confirm a Soybean Cyst Nematode Resistance Gene from PI 88788 on Linkage Group J**. *Crop Sci* 2004, **44**:936–941.
41. Vaghchhipawala Z, Bassüner R, Clayton K, Lewers K, Shoemaker R, Mackenzie S: **Modulations in Gene Expression and Mapping of Genes Associated with Cyst Nematode Infection of Soybean**. *Am Phytopathol Soc* 2001, **14**:42–54.
42. Concibido VC, Young ND, Lange DA, Denny RL, Danesh D, Orf JH: **Targeted comparative genome analysis and qualitative mapping of a major partial . resistance gene to the soybean cyst nematode**. *Theor Appl Genet* 1996, **93**:234–241.
43. Liang H, Yu Y, Wang S, Lian Y, Wang T, Wei Y, Gong P, Liu X, Fang X, Zhang M: **QTL Mapping of Isoflavone, Oil and Protein Contents in Soybean (Glycine max L. Merr.)**. *Agric Sci China* 2010, **9**:1108–1116.
44. Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerrière A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D: **APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy.** *Nat Genet* 2006, **38**:24–6.
45. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King M, Skuse D, Geschwind DH, Gilliam TC, et al.: **Strong Association of De Novo Copy Number Mutations with Autism**. *Science (80-)* 2010, **316**:445–449.
46. Simon-sanchez J, Scholz S, Matarin M del M, Fung H, Hernandez D, Gibbs JR, Britton A, Hardy J, Singleton A: **Genomewide SNP Assay Reveals Mutations Underlying Parkinson Disease**. *Hum Mutat* 2008, **29**:315–322.
47. Cook DE, Bayless AM, Wang K, Guo X, Song Q, Jiang J, Bent AF: **Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode.** *Plant Physiol* 2014, **165**:630–647.
48. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.
49. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo MA: **The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data**. 2010:1297–1303.
50. Liu Y, Khan SM, Wang J, Chen S, Rynge M, Wang J, Santos JVM dos, Valliyodan B, Merchant N, Nguyen HT, Xu D, Joshi T: **Large Scale NGS resequencing data analysis workflow for soybean germplasm using iPlant, XSEDE and SoyKB framework**. *Bioinformatics* 2014, **in press**.

51. **The Extreme Science and Engineering Discovery Environment (XSEDE)** [https://www.xsede.org/home]
52. Goff S a, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, Muir A, Merchant N, Lowry S, Mock S, Helmke M, Kubach A, Narro M, Hopkins N, Micklos D, Hilgert U, Gonzales M, Jordan C, Skidmore E, Dooley R, Cazes J, McLay R, Lu Z, Pasternak S, Koesterke L, Piel WH, et al.: **The iPlant Collaborative: Cyberinfrastructure for Plant Biology.** *Front Plant Sci* 2011, **2**(July):34.
53. Deelman E, Singh G, Su M, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman GB, Good J, Laity A, Jacob JC, Katz DS: **Pegasus : A framework for mapping complex scientific workflows onto distributed systems**. *Sci Program* 2005, **13**(January):219–237.
54. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S: **cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate.** *Nucleic Acids Res* 2012, **40**:e69.
55. Joshi T, Fitzpatrick MR, Chen S, Liu Y, Zhang H, Endacott RZ, Gaudiello EC, Stacey G, Nguyen HT, Xu D: **Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding.** *Nucleic Acids Res* 2014, **42**(Database issue):D1245–52.
56. Joshi T, Patil K, Fitzpatrick MR, Franklin LD, Yao Q, Cook JR, Wang Z, Libault M, Brechenmacher L, Valliyodan B, Wu X, Cheng J, Stacey G, Nguyen HT, Xu D: **Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics.** *BMC Genomics* 2012, **13 Suppl 1**(Suppl 1):S15.
57. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3**. *Landes Biosci* 2012, **6**:80–92.
58. **ANALYSIS TOOLKIT FOR THE AGRICULTURAL COMMUNITY (agriGO)** [http://bioinfo.cau.edu.cn/agriGO/analysis.php]
59. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731–9.
60. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156–8.

Figure and Table Legends

Table 1. Basic description about all the Brazilian and U.S. soybeans accessions used in this study

Table 2. Number of unique SNPs for each Brazilian cultivars.

Table 3. Summary of regions under positive selection process with F_{ST} and nucleotide diversity ($\theta\pi$) values.

Figure 1. Summary of the main modification caused by SNPs and Indels.

Figure 2. Population structure analysis of the 28 Brazilian soybean cultivars. **(a)** Neighbor-joining phylogenetic tree generated for the 28 Brazilian soybean accessions. **(b)** Bayesian clustering (FastStructure, K = 3) for the 28 Brazilian soybean cultivars.

Figure 3. Two regions between 3.01-3.09 Mb (a) and 5.53-5.92 Mb (b) on chromosome 17 under positive selection.

Figure 4. Copy Number Variations (CNVs) detected on chromosome 16 for oldest and latest Brazilian cultivars.

List of Supplementary Accessions

Supplementary Table 1. Sequencing information of the Brazilian lines.

Supplementary Table 2. Variant rate details of the Brazilian soybeans accessions.

Supplementary Table 3. Number of SNPs identified on coding regions in Brazilian soybeans

Supplementary Table 4. Number of genes with allelic variation observed in Brazilian lines.

Supplementary Table 5. Summary of the most relevant results from GO enrichment analysis.

Supplementary Figure 1. Number of homozygous/heterozygous SNPs (a) and Indels (b) for each Brazilian soybean used in this study.

Supplementary Figure 2. Copy Number Variation (CNV) for each Brazilian lines used in this study.

Supplementary Figure 3. Copy Number Variations (CNVs) detected in Brazilian cultivars on chromosome 6, 7, 8, 9,13, 15 and 17.

Supplementary Figure 4. Copy number variations (CNVs) observed between Brazilian and U.S. accessions.

Table 1: Basic description about all the Brazilian and U.S. soybeans accessions used in this study.

Access Name	Origin	Decade	Maturity Group
Anta 82	Brazil	2001-2010	7.2
BR 16	Brazil	1991-2000	6.4
BRS 232	Brazil	2001-2010	6.9
BRS 284	Brazil	2001-2010	6.3
BRS 360RR	Brazil	2011-2020	6.2
BRS 361	Brazil	2011-2020	7.4
BRS Sambaíba	Brazil	2001-2010	9.3
BRS Valiosa RR	Brazil	2001-2010	8.1
BRSGO 8360	Brazil	2001-2010	8.3
BRSGO 8660	Brazil	2001-2010	8.6
BRSGO Chapadões	Brazil	2001-2010	8.6
BRSMG 850 GRR	Brazil	2001-2010	8.2
BRSMT Pintado	Brazil	1991-2000	8.4
BRSMT Uirapuru	Brazil	1991-2000	9.0
CD 201	Brazil	1991-2000	6.6
Doko	Brazil	1981-1990	9.0
Embrapa 48	Brazil	1991-2000	6.8
EMGOPA 301	Brazil	1981-1990	Not Available
FT Abyara	Brazil	1981-1990	7.3-7.6
FT Cristalina	Brazil	1981-1990	7.6
IAC 8	Brazil	1971-1980	Not Available
IAS 5	Brazil	1971-1980	6.4
MG/BR46	Brazil	1991-2000	8.1
NA 5909 RG	Brazil	2001-2010	6.2
P98Y11	Brazil	2001-2010	8.1
Paraná	Brazil	1971-1980	Not Available
Santa Rosa	Brazil	1961-1970	Not Available
V MAX RR	Brazil	2001-2010	6.2
US-01	U.S.	1981-1990	V
US-02	U.S.	2011-2014	IV
US-03	U.S.	1971-1980	V
US-04	U.S.	2011-2014	IV
US-05	U.S.	1991-2000	V
US-06	U.S.	2001-2010	III
US-07	U.S.	1991-2000	III
US-08	U.S.	1981-1990	III
US-09	U.S.	1981-1990	III
US-10	U.S.	1971-1980	II
US-11	U.S.	1951-1960	VII
US-12	U.S.	1991-2000	III
US-13	U.S.	1991-2000	III
US-14	U.S.	1991-2000	III
US-15	U.S.	2001-2010	IV
US-16	U.S.	2001-2010	VIII
US-17	U.S.	1971-1980	V
US-18	U.S.	1991-2000	VII
US-19	U.S.	1991-2000	IV

Table 2. Number of unique SNPs for each Brazilian cultivars

Accession name	Start codon	Stop codon	Syn cds	Non-syn cds	Intron	Splice site	Other	Total
Anta 82	9	89	1,447	85	178	7	1,771	3,586
BR 16	11	5	121	159	359	17	6,364	7,036
BRS/GO 8360	13	4	127	158	405	18	4,603	5,328
BRS/GO 8660	22	15	265	356	1,091	38	18,601	20,388
BRS/GO Chapadões	104	43	1,130	1,384	3,796	152	67,705	74,314
BRS 232	9	1	83	123	190	9	3,236	3,651
BRS 284	33	16	418	592	1,967	77	59,176	62,279
BRS 360RR	7	1	95	124	315	16	3,173	3,731
BRS 361	17	7	144	231	754	21	9,604	10,778
BRS Sambaíba	45	25	511	641	1,591	60	28,938	31,811
BRS Valiosa RR	159	0	1	8	21	1	154	344
BRSMG_850GRR	0	0	1	8	18	2	289	318
BRSMT Pintado	12	1	75	86	156	13	2,773	3,116
BRSMT Uirapuru	9	7	134	172	398	17	9,925	10,662
CD 201	28	2	182	235	647	19	9,937	11,050
Conquista	4	2	15	25	80	2	1,358	1,486
Doko	58	25	655	808	2,006	82	39,192	42,826
Embrapa 48	1	0	60	70	103	7	1,641	1,882
Emgopa 301	20	17	234	333	765	41	11,180	12,590
FT Abyara	86	23	705	884	2,396	75	3,2278	36,447
FT Cristalina	2	0	7	11	26	3	409	458
IAC 8	20	30	379	507	1,440	61	38,888	41,325
IAS 5	26	3	251	260	503	32	7,843	8,918
NA 5909 RG	34	10	383	490	1,334	49	20,391	22,691
P98Y11	22	9	196	285	519	33	17,526	18,590
Paraná	1	2	60	108	220	13	6,431	6,835
Santa Rosa	86	42	958	1,416	4,273	162	89,168	96,105
VMAX RR	6	4	98	107	205	18	2,777	3,215
Total	844	383	8,735	9,666	25,756	1,045	495,331	541,760

*****Non coding region: correspond to 3'UTR, 5' UTR and intergenic region; Syn cds: synonymous SNP inside coding region; Non-syn cds: non-synonymous SNP inside coding region, Other: all the remaining genome regions

Table 3. Summary of regions under positive selection process with F_{ST} and nucleotide diversity (θπ) values.

Chromosome	Start	End	Number of SNPs	θπ (oldest cultivars)	θπ (latest cultivars)	F_{ST}
07	40,100,001	40,110,000	41	0.00219	0.00000	0.7071
	40,110,001	40,120,000	12	0.00064	0.00000	0.7071
	40,140,001	40,150,000	26	0.00139	0.00000	0.7071
	40,150,001	40,160,000	31	0.00165	0.00013	0.7071
	40,160,001	40,170,000	36	0.00192	0.00006	0.7071
	40,630,001	40,640,000	21	0.00112	0.00000	0.7071
15	2,950,001	2,960,000	35	0.00187	0.00014	0.7071
	2,960,001	2,970,000	16	0.00085	0.00000	0.7071
	3,010,000	3,020,000	17	0.00060	0.00000	0.8695
	3,030,001	3,040,000	23	0.00082	0.00000	0.8194
	3,040,001	3,050,000	41	0.00146	0.00002	0.8695
	3,050,001	3,060,000	13	0.00046	0.00000	0.7486
	5,560,001	5,570,000	76	0.00279	0.00000	0.8620
	5,570,001	5,580,000	31	0.00110	0.00000	0.8695
	5,580,001	5,590,000	26	0.00092	0.00000	0.8695
	5,610,001	5,620,000	22	0.00078	0.00000	0.8275
	5,620,001	5,630,000	34	0.00121	0.00000	0.8695
	5,660,001	5,670,000	39	0.00140	0.00000	0.8677
	5,670,001	5,680,000	26	0.00092	0.00000	0.8695
	5,680,001	5,690,000	35	0.00128	0.00000	0.8383
	5,710,001	5,720,000	28	0.00100	0.00003	0.8695
	5,730,001	5,740,000	20	0.00070	0.00000	0.8321
	5,740,001	5,750,000	45	0.00160	0.00004	0.8695
17	5,750,001	5,760,000	26	0.00094	0.00000	0.8572
	5,760,001	5,770,000	74	0.00263	0.00000	0.8695
	5,770,001	5,780,000	24	0.00088	0.00001	0.8636
	5,780,001	5,790,000	39	0.00139	0.00000	0.8676
	5,790,001	5,800,000	25	0.00089	0.00000	0.8695
	5,800,001	5,810,000	63	0.00224	0.00000	0.8671
	5,810,001	5,820,000	50	0.00178	0.00000	0.8695
	5,820,001	5,830,000	48	0.00171	0.00000	0.8695
	5,830,001	5,840,000	48	0.00171	0.00003	0.8679
	5,840,001	5,850,000	27	0.00096	0.00000	0.8695
	5,850,001	5,860,000	24	0.00085	0.00007	0.8695
	5,860,001	5,870,000	69	0.00249	0.00010	0.8664
	5,870,001	5,880,000	32	0.00114	0.00000	0.8695
	5,880,001	5,890,000	66	0.00238	0.00000	0.8663
	5,890,001	5,900,000	76	0.00270	0.00003	0.8447
	5,900,001	5,910,000	58	0.00206	0.00000	0.8695
	5,910,001	5,920,000	14	0.00050	0.00007	0.8050
18	2,190,001	2,200,000	107	0.00571	0.00010	0.7032

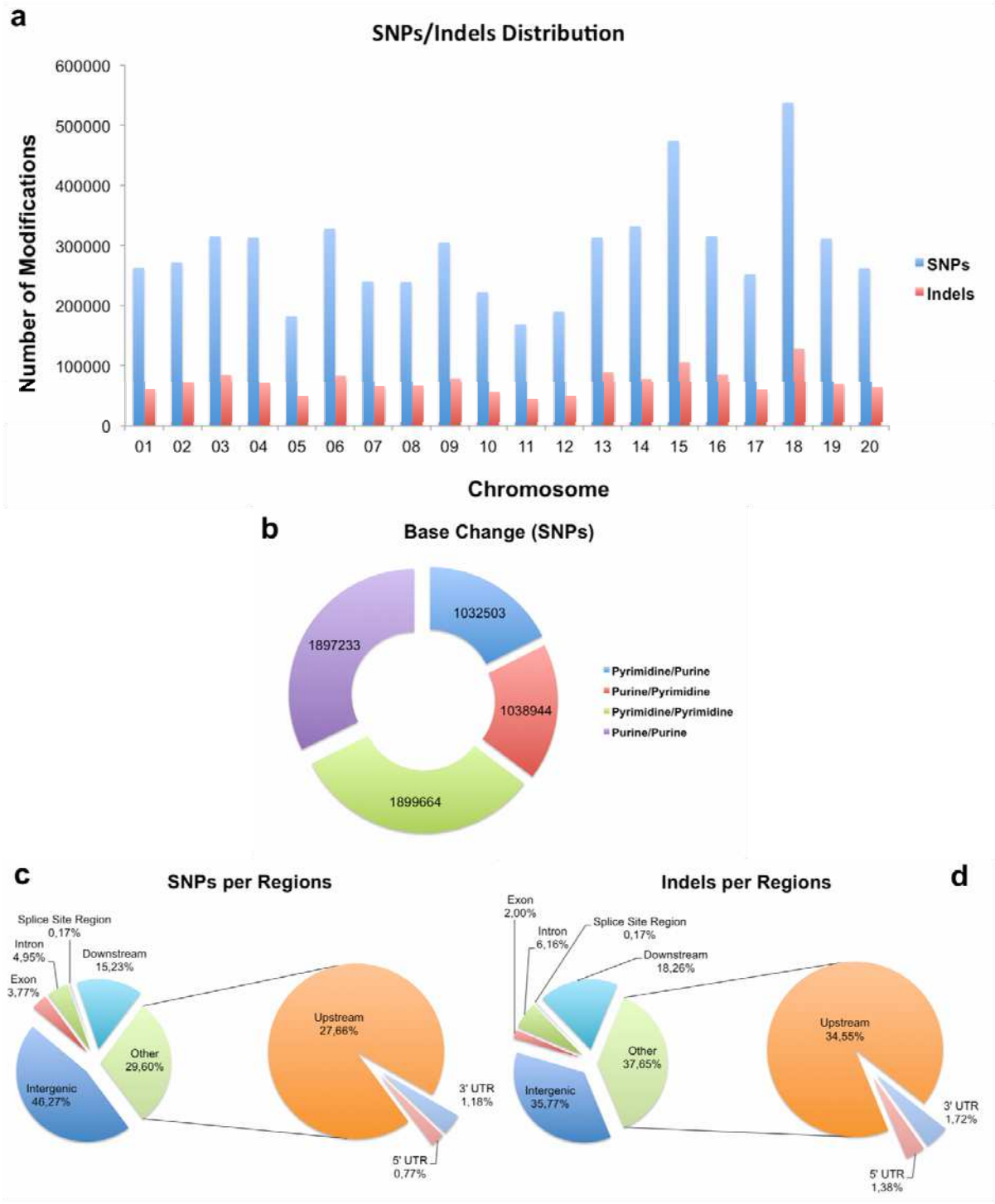


Figure 1. Summary of the main modification caused by SNPs and Indels. (**a**) SNPs (blue) and Indels (red) distribution by the 20 soybean chromosomes. (**b**) Number of transition/transversion mutations: Pyrimidine/Purine (blue), Purine/Pyrimidine (red), Pyrimidine/Pyrimidine (green) and Purine/Purine (purple). (**c**) Percentage of SNPs per region of the soybean genome. (**d**) Percentage of Indels per region of the soybean genome.

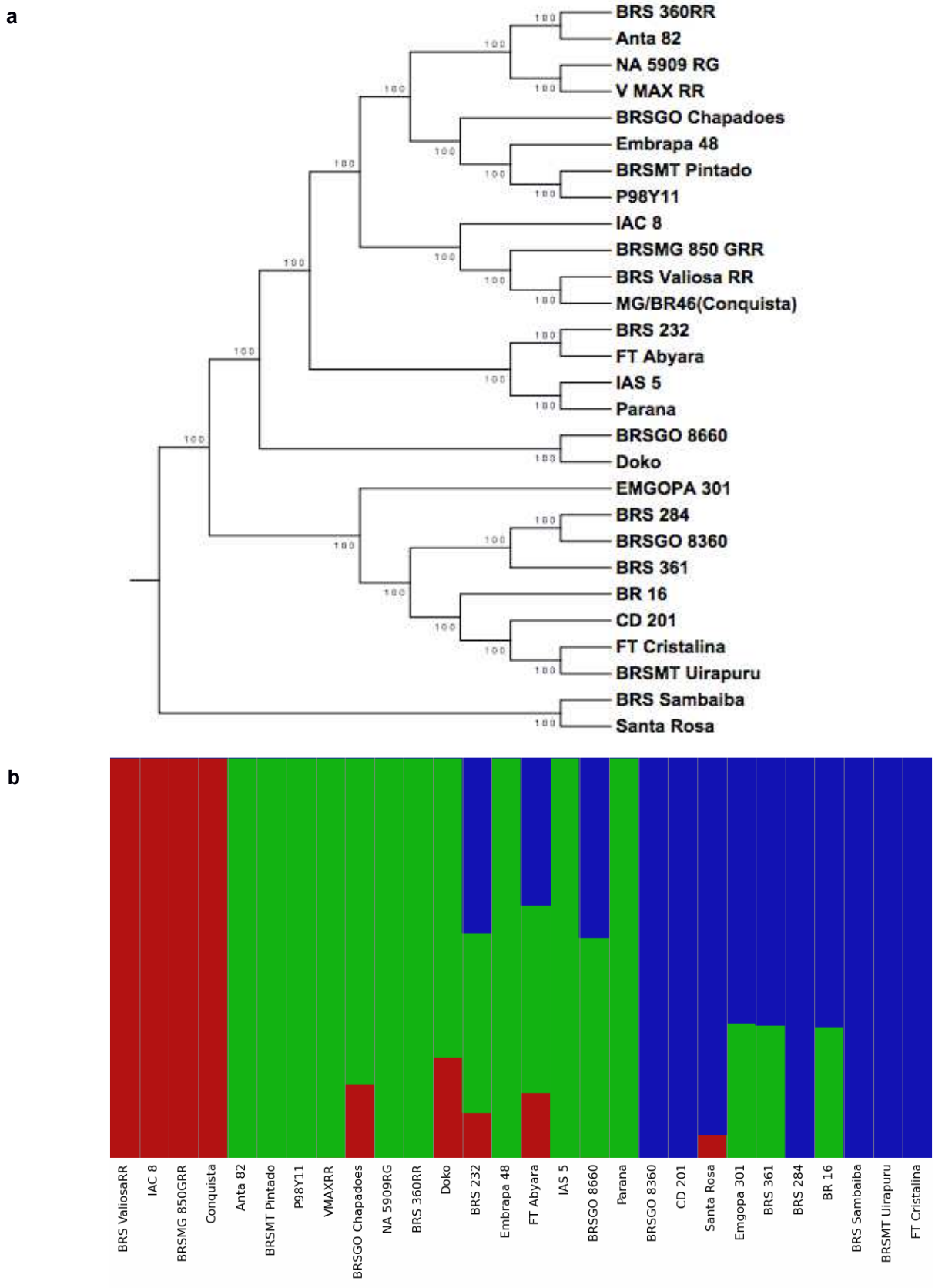


Figure 2. Population structure analysis of the 28 Brazilian soybean cultivars. **(a)** Neighbor-joining phylogenetic tree generated for the 28 Brazilian soybean accessions. **(b)** Bayesian clustering (FastStructure, *K* = 3) for the 28 Brazilian soybean cultivars.

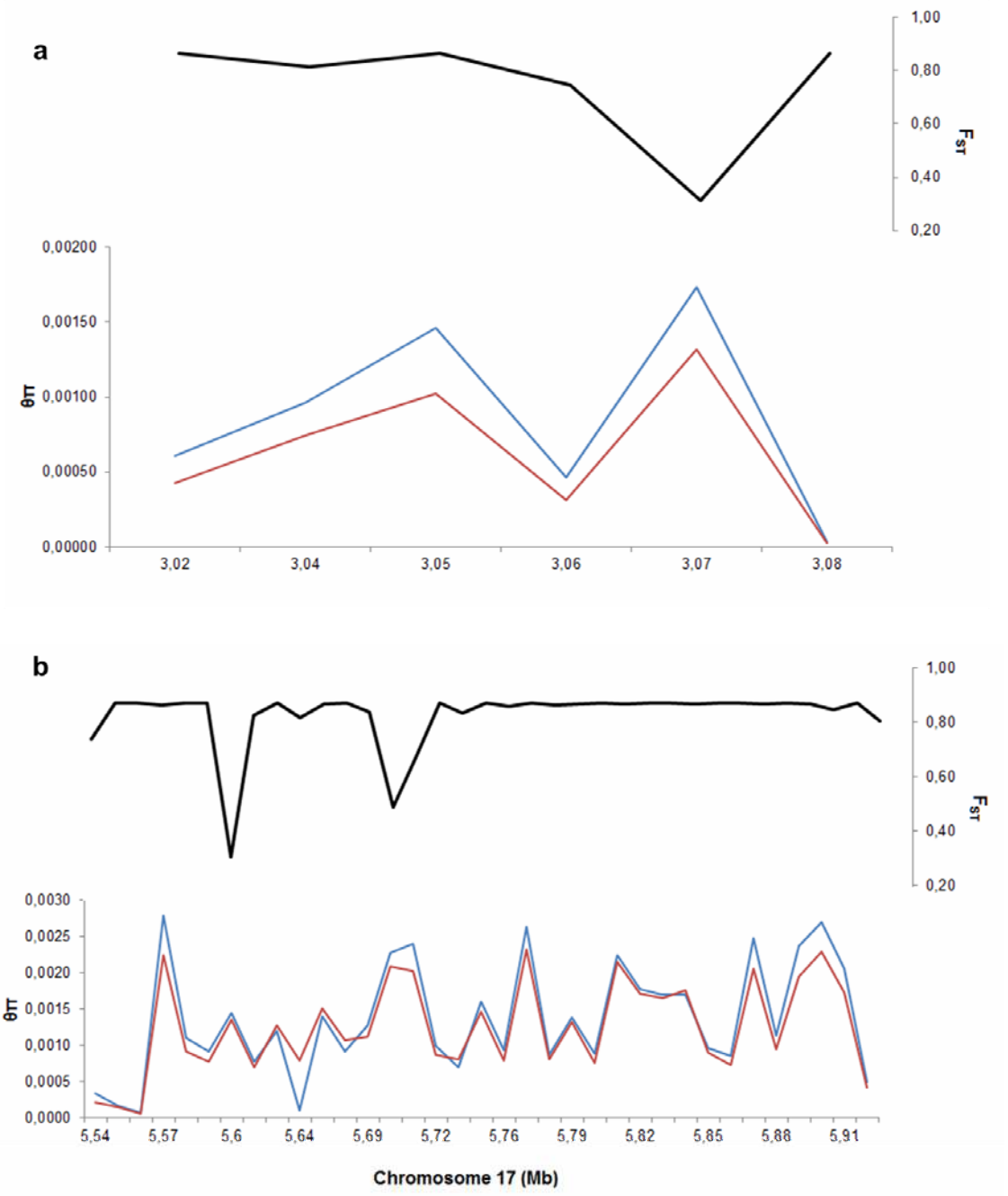


Figure 3. Two regions between 3.01-3.09 Mb (a) and 5.53-5.92 Mb (b) on chromosome 17 under positive selection. The red line corresponds to the nucleotide diversity of latest cultivars and the blue line to the oldest cultivars. The black line is the F_{ST} values between oldest and latest cultivars.

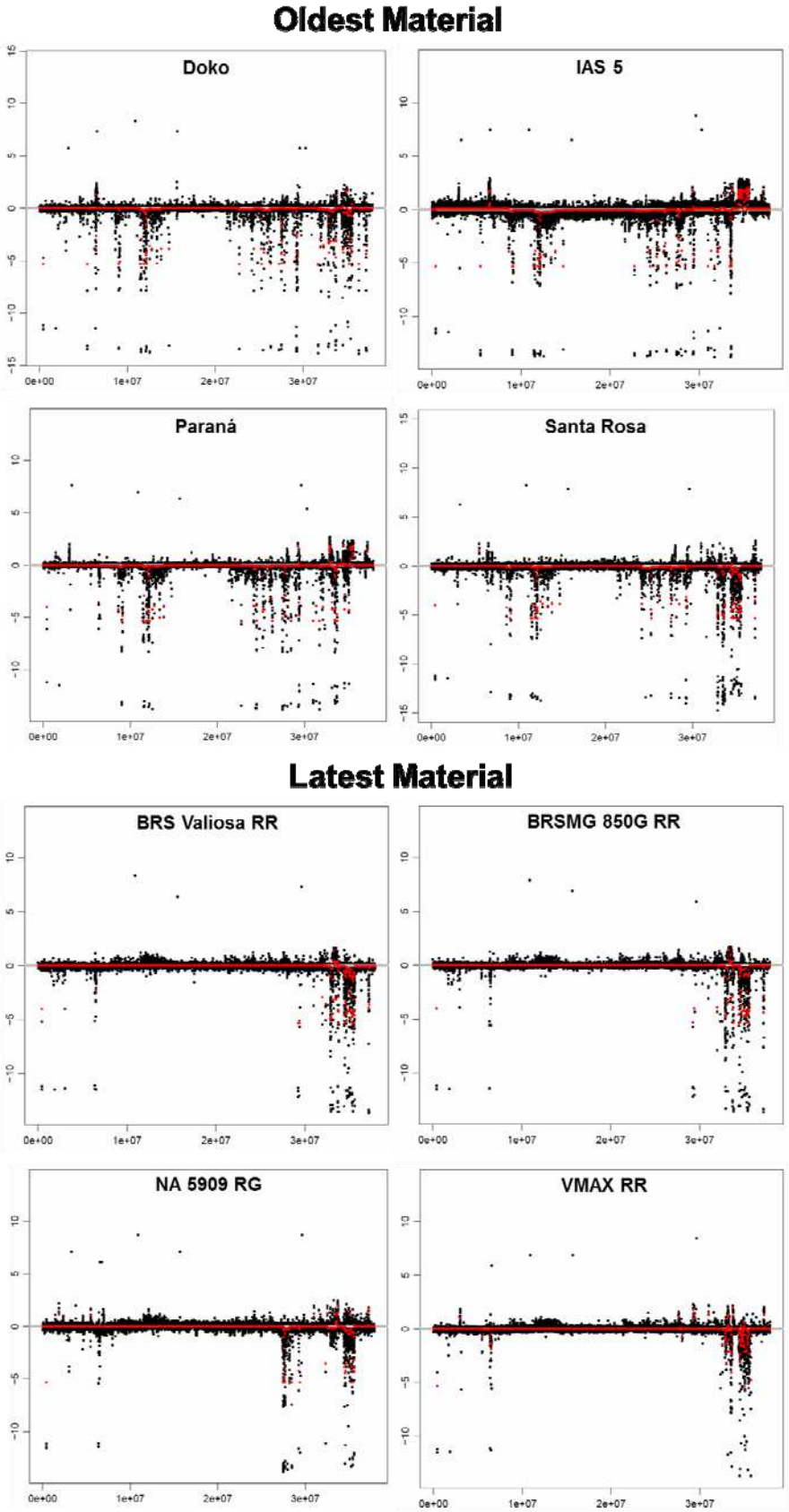


Figure 4. Copy Number Variations (CNVs) detected on chromosome 16 for oldest and latest Brazilian cultivars The x-axis represents the genomic position and y-axis the log-ratio of the read counts. The red dots are the copy number call of each segment.

Supplementary Table 1. Sequencing information of the Brazilian lines.

Accession Name	Number of reads	Number of mapped reads	Genome coverage	Mean depth
Anta 82	195,886,570	180,622,083	0.9221	13.2579
BR 16	196,397,054	186,322,940	0.9487	14.7765
BRS 232	122,368,531	116,984,169	0.9560	9.1884
BRS 284	171,461,982	163,470,705	0.9534	12.6584
BRS 360RR	201,402,004	192,681,470	0.9567	15.3996
BRS 361	191,425,924	183,136,909	0.9567	14.6088
BRS Sambaíba	192,188,713	181,426,603	0.9440	14.7690
BRS Valiosa RR	156,259,988	148,106,406	0.9478	11.4630
BRSGO 8360	158,327,168	148,939,413	0.9407	11.1166
BRSGO 8660	142,045,602	133,525,390	0.9400	10.5473
BRSGO Chapadões	245,514,627	222,295,561	0.9054	17.6885
BRSMG 850 GRR	217,891,058	206,849,339	0.9493	16.4315
BRSMT Pintado	191,564,772	181,730,846	0.9487	14.4552
BRSMT Uirapuru	198,605,106	186,387,017	0.9385	15.2175
CD 201	221,562,422	210,956,658	0.9521	17.2033
Doko	247,331,922	234,203,818	0.9469	19.2339
Embrapa 48	211,246,973	195,673,675	0.9263	15.9426
EMGOPA 301	159,411,439	150,539,582	0.9443	12.1211
FT Abyara	185,414,289	175,454,862	0.9463	13.9882
FT Cristalina	223,191,658	213,329,190	0.9558	17.6043
IAC 8	204,515,975	194,693,512	0.9520	15.7266
IAS 5	150,707,921	133,763,689	0.8876	10.8346
MG/BR46	272,124,638	258,659,654	0.9505	21.4019
NA 5909 RG	189,097,657	179,353,466	0.9485	13.8281
P98Y11	149,136,838	141,851,602	0.9512	11.2006
Paraná	317,943,922	301,620,197	0.9487	23.9907
Santa Rosa	166,171,639	157,805,234	0.9497	12.6687
V MAX RR	221,332,482	207,971,094	0.9396	16.3037

Supplementary Table 2. Variant rate details of the Brazilian soybeans accessions.

Chromosome	Length	Number of SNPs	Variants rate
1	56,831,624	263,927	215
2	48,577,505	273,274	177
3	45,779,781	317,779	144
4	52,389,146	313,989	166
5	42,234,498	182,558	231
6	51,416,486	330,362	155
7	44,630,646	241,445	184
8	47,837,940	240,096	199
9	50,189,764	306,025	164
10	51,566,898	223,063	231
11	34,766,867	169,064	205
12	40,091,314	190,138	210
13	45,874,162	315,257	145
14	49,042,192	333,541	147
15	51,756,343	477,983	108
16	37,887,014	318,763	118
17	41,641,366	253,241	164
18	58,018,742	541,951	107
19	50,746,916	312,722	162
20	47,904,181	263,166	182
Total	**949,183,385**	**5,868,344**	**161**

Supplementary Table 3. Number of SNPs identified in coding regions of Brazilian soybeans cultivars

Cultivars	Modifications							Total
	Non-syn cds	Start		Stop		Splice Site		
		G.	L.	G.	L.	A.	D.	
All	551	40	8	2	4	0	4	609
Anta 82	19,262	1,212	54	404	86	144	128	21,290
BR 16	30,499	1,904	76	610	130	205	182	33,606
BRS 232	29,573	1,785	74	621	132	191	184	32,560
BRS 284	20,802	1,219	45	434	98	137	121	22,856
BRS 360 RR	24,112	1,481	70	489	103	179	168	26,602
BRS 361	22,546	1,321	62	478	103	165	128	24,803
BRS Chapadões	29,089	1,707	79	583	112	223	177	31,970
BRS Sambaíba	31,549	1,936	81	672	136	128	194	34,696
BRS Valiosa RR	30,888	1,871	96	640	127	211	181	34,014
BRSGO 8360	21,098	1,219	47	475	98	140	124	23,201
BRSGO 8660	31,253	1,926	79	677	139	226	202	34,502
BRSMG 850G RR	30,119	1,857	92	613	127	210	172	33,190
BRSMT Pintado	28,499	1,748	78	596	121	195	179	31,416
BRSMT Uirapuru	31,069	1,925	67	664	155	212	221	34,313
CD 201	28,222	1,741	77	588	122	228	185	31,163
Conquista	30,221	1,878	96	631	124	209	172	33,331
Doko	31,969	2,029	92	663	138	225	186	35,302
Embrapa 48	26,656	1,564	68	522	113	182	162	29,267
Emgopa 301	28,117	1,755	69	630	121	197	197	31,086
FT Abyara	29,076	1,81	81	611	145	215	170	32,108
FT Cristalina	30,845	1,935	70	661	148	219	209	34,087
IAC 8	29,473	1,797	84	644	122	192	157	32,469
IAS 5	27,797	1,758	65	585	126	199	183	30,713
NA 5909 RG	21,521	1,271	64	434	81	172	141	23,684
P98Y11	31,065	1,881	82	670	131	229	196	34,254
Paraná	26,256	1,705	66	555	124	185	164	29,055
Santa Rosa	33,872	2,058	100	748	152	236	218	37,384
VMAX RR	20,699	1,254	65	437	79	162	137	22,833

All: SNP present in all Brazilian cultivars compared to reference genome; Non-syn cds: non-synonymous SNP inside coding region, Start G.: A variant in 5'UTR region produces a three base sequence that can be a START codon; Start L.: Variant causes start codon to be mutated into a non-start codon; Stop G.: Variant causes a STOP codon; Stop L.: Variant causes stop codon to be mutated into a non-stop codon; Splice Site A.: The variant hits a splice acceptor site; Splice Site D.: The variant hits a Splice donor site.

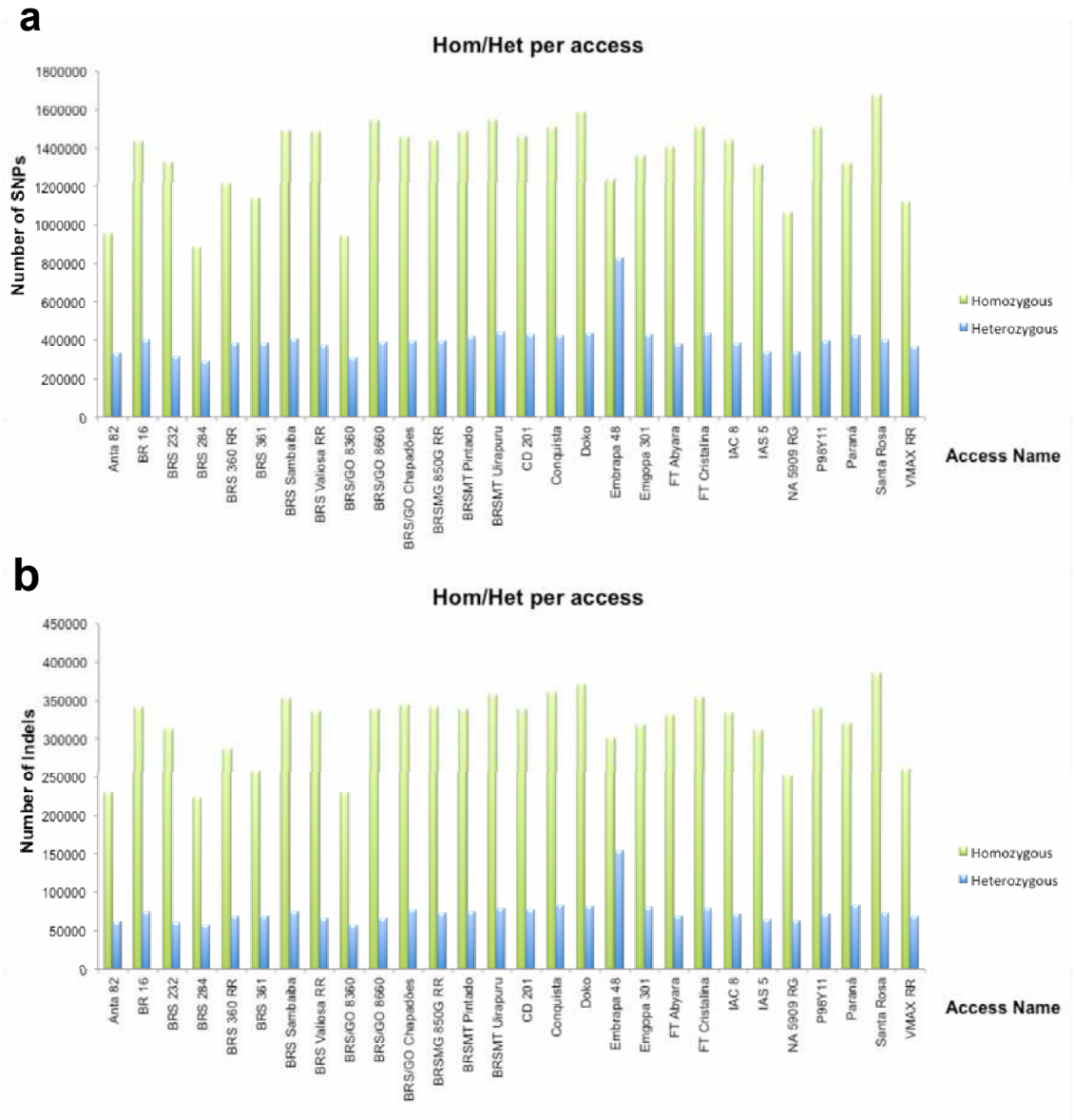
Supplementary Table 4. Number of genes with allelic variation observed in Brazilian lines.

Cultivars	Modifications							Total
	Non-syn cds	Start Codon		Stop Codon		Splice Site		
		G.	L.	G.	L.	D.	A.	
All	334	35	7	2	6	5	0	369
Anta 82	10,638	1,403	68	529	137	190	205	11,515
BR 16	7,642	989	57	376	97	138	152	8,287
BRS 232	11,644	1,566	78	546	140	189	210	12,632
BRS 284	11,035	1,463	77	564	143	192	197	11,954
BRS 360 RR	7,673	990	47	392	108	131	141	8,301
BRS 361	9,41	1,231	72	456	114	175	182	10,186
BRS Chapadões	8,066	1,006	68	405	89	147	168	8,748
BRS Sambaíba	8,414	1,089	63	430	114	138	167	9,084
BRS Valiosa RR	11,124	1,422	81	544	123	184	231	12,022
BRSGO 8360	8,018	1,029	67	398	91	151	173	8,688
BRSGO 8660	11,892	1,555	83	601	146	201	134	12,895
BRSMG 850G RR	11,531	1,53	98	579	138	189	214	12,508
BRSMT Pintado	7,695	1,006	50	413	110	131	146	8,314
BRSMT Uirapuru	11,581	1,519	83	603	141	203	230	12,542
CD 201	11,828	1,544	80	596	148	209	225	12,804
Conquista	12,046	1,656	91	600	149	193	227	13,061
Doko	10,253	1,388	69	506	134	175	188	11,119
Embrapa 48	11,261	1,522	94	562	138	178	212	12,220
Emgopa 301	12,84	1,702	101	665	162	223	238	13,907
FT Abyara	11,128	1,442	80	541	132	187	200	12,055
FT Cristalina	11,608	1,553	69	595	165	227	219	12,569
IAC 8	10,994	1,427	78	547	132	192	233	11,921
IAS 5	11,386	1,542	98	577	135	180	211	12,378
NA 5909 RG	10,223	1,307	70	483	124	167	189	11,026
P98Y11	10,614	1,446	72	570	133	200	200	11,534
Paraná	11,032	1,473	83	557	155	178	218	11,945
Santa Rosa	11,588	1,530	71	587	159	216	224	12,532
VMAX RR	11,066	1,449	86	576	133	165	198	11,977

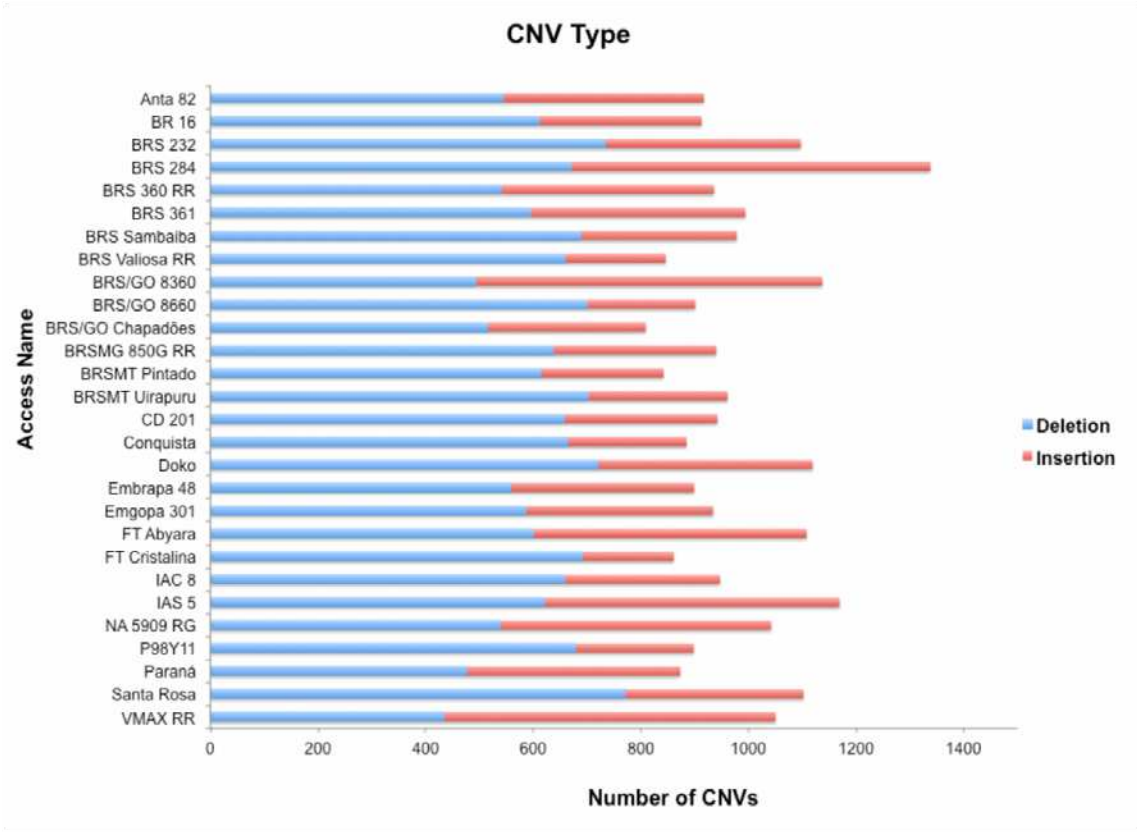
All: gene allelic modification present in all Brazilian cultivars compared to reference genome; Non-syn cds: non-synonymous SNP inside coding region, Start G.: A variant in 5'UTR region produces a three base sequence that can be a START codon; Start L.: Variant causes start codon to be mutated into a non-start codon; Stop G.: Variant causes a STOP codon; Stop L.: Variant causes stop codon to be mutated into a non-stop codon; Splice Site A.: The variant hits a splice acceptor site; Splice Site D.: The variant hits a Splice donor site.

Supplementary Table 5. Summary of the most relevant results from GO enrichment analysis.

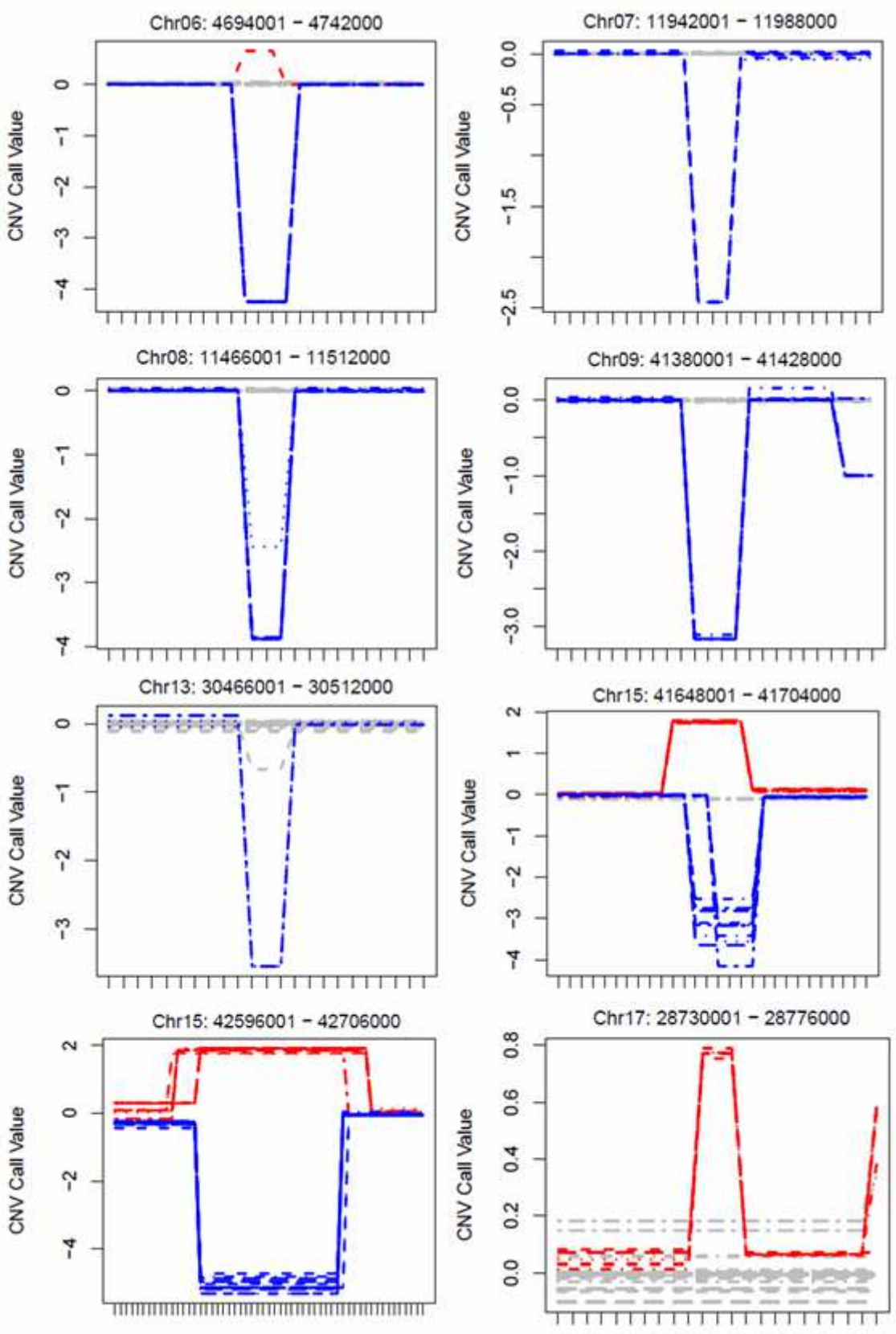
Description	Number of genes	Genes
Generation of precursor metabolites and energy	34	Glyma.01G058600, Glyma.01G076000, Glyma.01G091900, Glyma.01G095900, Glyma.01G153500, Glyma.01G201600, Glyma.04G095000, Glyma.04G204000, Glyma.05G073600, Glyma.06G217900, Glyma.07G143800, Glyma.08G281300, Glyma.09G171300, Glyma.09G178600, Glyma.10G226400, Glyma.11G114700, Glyma.12G056400, Glyma.13G088500, Glyma.13G155500, Glyma.14G149900, Glyma.15G114600, Glyma.15G188400, Glyma.15G239000, Glyma.15G238900, Glyma.18G078500, Glyma.18G155300, Glyma.18G155400, Glyma.18G203700, Glyma.19G051900, Glyma.19G053400, Glyma.19G054200, Glyma.19G081000, Glyma.19G083500, Glyma.19G109600
DNA-dependent transcription, elongation	34	Glyma.01G058600, Glyma.01G076000, Glyma.01G091900, Glyma.01G095900, Glyma.01G153500, Glyma.01G201600, Glyma.04G095000, Glyma.04G204000, Glyma.05G073600, Glyma.06G217900, Glyma.07G143800, Glyma.08G281300, Glyma.09G171300, Glyma.09G178600, Glyma.10G226400, Glyma.11G114700, Glyma.12G056400, Glyma.13G088500, Glyma.13G155500, Glyma.14G149900, Glyma.15G114600, Glyma.15G188400, Glyma.15G239000, Glyma.15G238900, Glyma.18G078500, Glyma.18G155300, Glyma.18G155400, Glyma.18G203700, Glyma.19G051900, Glyma.19G053400, Glyma.19G054200, Glyma.19G081000, Glyma.19G083500, Glyma.19G109600
Photosynthesis	38	Glyma.01G058600, Glyma.01G076000, Glyma.01G091900, Glyma.01G095900, Glyma.01G153500, Glyma.01G201600, Glyma.04G095000, Glyma.04G204000, Glyma.05G073600, Glyma.06G217900, Glyma.06G228400, Glyma.07G143800, Glyma.08G281300, Glyma.09G171300, Glyma.09G178600, Glyma.10G226400, Glyma.11G081100, Glyma.11G114700, Glyma.12G056400, Glyma.13G088500, Glyma.13G155500, Glyma.14G149900, Glyma.15G114600, Glyma.15G188400, Glyma.15G239000, Glyma.15G238900, Glyma.15G246000, Glyma.18G078500, Glyma.18G155300, Glyma.18G155400, Glyma.18G203700, Glyma.18G262700, Glyma.19G051900, Glyma.19G053400, Glyma.19G054200, Glyma.19G081000, Glyma.19G083500, Glyma.19G109600
Photosynthesis, light reaction	19	Glyma.01G058600, Glyma.01G095900, Glyma.01G153500, Glyma.04G095000, Glyma.05G073600, Glyma.06G217900, Glyma.07G143800, Glyma.07G201300, Glyma.08G281300, Glyma.11G114700, Glyma.12G056400, Glyma.13G088500, Glyma.15G114600, Glyma.18G155300, Glyma.18G155400, Glyma.19G053400, Glyma.19G081000, Glyma.19G083500, Glyma.19G109600
ATP synthesis coupled electron transport	6	Glyma.01G101600, Glyma.06G228400, Glyma.10G068800, Glyma.11G081100, Glyma.15G246000, Glyma.18G155400
Photosynthetic electron transport in photosystem II	6	Glyma.01G153500, Glyma.04G095000, Glyma.05G073600, Glyma.06G217900, Glyma.11G114700, Glyma.18G262700
Cellular respiration	7	Glyma.06G228400, Glyma.11G081100, Glyma.12G056400, Glyma.15G246000, Glyma.18G262700, Glyma.19G081000, Glyma.19G083500



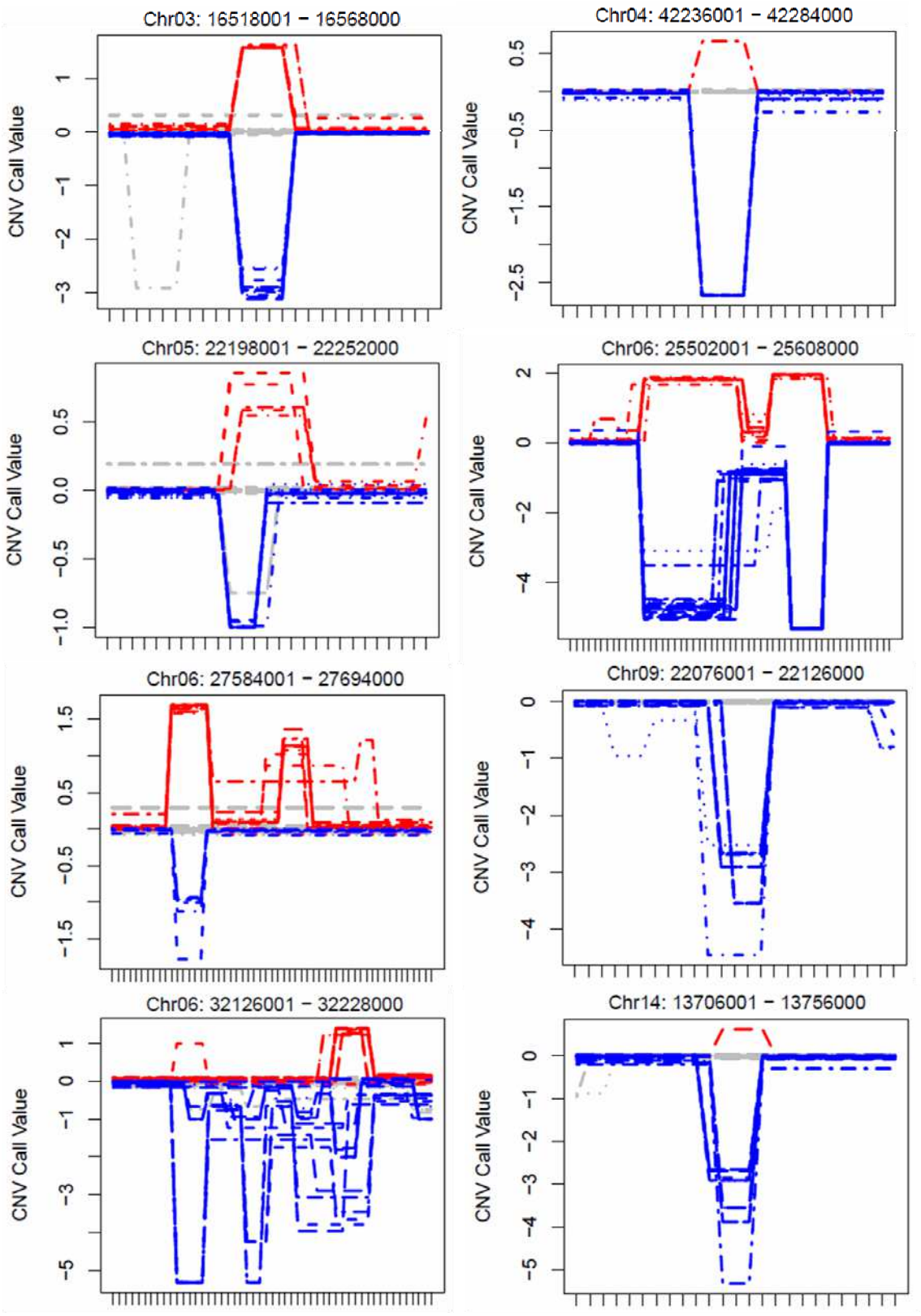
Supplementary Figure 1. Number of homozygous/heterozygous SNPs **(a)** and Indels **(b)** for each Brazilian soybean used in this study.



Supplementary Figure 2. Copy Number Variation (CNV) for each Brazilian lines used in this study.



Supplementary Figure 3. Copy Number Variations (CNVs) detected in Brazilian cultivars on chromosome 6, 7, 8, 9,13, 15 and 17. The x-axis represents the genomic position and the y-axis the CNV call produced by the segmentation algorithm. The blue lines are deleted fragments detected in these regions.



Supplementary Figure 4. Copy number variations (CNVs) observed between Brazilian and U.S. accessions. The x-axis represents the genomic position and the y-axis the CNV call produced by the segmentation algorithm. The red/blue lines are inserted/deleted fragments detected in these regions.

ARTIGO II - RESEQUENCING STRATEGIES FOR GENOME-WIDE ASSOCIATION STUDIES FOR IMPORTANT BIOTIC STRESSES FOR SOYBEAN

Background

Worldwide, soybean [*Glycine max* (L) Merrill] is one of the most important crops due to the major importance in human food and biofuel production. Therefore, a great number of studies to better understand the variation of the soybean genome are been developed. There are a large number of biotic factors economically important in soybean crops, responsible for meaningful decreases of production. Among the available options to decrease the influence of biotic stress in soybean crops, cultivars with resistance genes against a specific pathogen appear to be the best choice. Thus genome-wide association studies (GWAS) may offer a great opportunity to identify and localize resistance genes, being an important tool for breeding programs. In this study, we perform a GWAS to find important SNPs and CNVs related to the resistance of three important soybean diseases: soybean cyst nematode (SCN) (*Heterodera glycines*) soybean root-knot nematode (RKN) (*Meloidogyne incognita*) and soybean stem canker (SSC) (*Diaporthe phaseolorum f. sp. meridionalis*).

Results

A total of 10,079 SNPs were identified related to resistance against the three soybeans diseases analyzed in this study. For SCN, a total of 3,615 SNPs were identified associated with resistance to race 1 and 3 in *Rhg* genes and QTLs spreaded on chromosomes 1, 7, 8, 9, 10, 11, 17, 18, 19 and 20. Moreover, we found 4,461 and 2,014 SNPs related to resistance against RKN and SSC, respectively. Chromosome 10 had all the allelic variation associated with RKN resistance, whereas *Rdm?* putative region on chromosome 14 had all the allelic variation associated with SSC resistance. Finally, a large number of CNVs were identified and could also be related to defense mechanisms of the plant.

Conclusion

This study provides promising results to breeding programs for diseases resistance and can be directly used in marker-assisted programs to select allelic variations responsible for the main defense mechanism of these three biotic factors. Our results open a new horizon for genotyping and the identification of accessions with resistance against SCN, RKN and SSC. However, a validation process of the SNPs will be necessary to confirm the results obtained in this study. In addition, a depth

study of the detected CNVs should be made due the possibility of large influence of these structural modifications in defense mechanism against pathogens actions.

Keywords: soybean, resistance mechanism, resequencing, allelic variation, GWAS, CNVs

Background

Soybean [*Glycine max* (L) Merrill] is one of the most important worldwide crops. It is estimated that the wild soybean (*Glycine soja*) was domesticated to cultivated soybean around of 7000–9000 years ago in Asia [1]. Due its major importance to human food and biofuel production a large number of studies to better understand the variation in soybean genome are been developed.

One of the biggest barriers to increase the soybean production and seed quality are the large number of biotic factors that affect soybean crops. In 2009, it was estimated losses around 484,451,000 bushels of soybeans caused by diseases in the United States [2]. These results highlight the importance of studies to develop cultivars with better performance under the influence of the main soybean biotic factors.

There are several strategies commonly used to decrease the influence of the biotic factors on soybean crop. Among the available options cultivars with resistance genes against a specific pathogen appear to be the best choice. A large number of soybean resistance genes for different diseases has been identified and mapped on the 20 years. Such knowledge associated with the identification and localization of new major genes responsible for the plant resistance to pathogen is the best option for breeding programs to develop cultivars with great performance under a pathogen attack. Thus genome-wide association studies (GWAS) may offer a great opportunity to identify and localize these resistance genes and shows as an important tool for breeding programs.

The advent of the new platform for large-scale sequencing allowed that a great number of variations could be identified and used in GWAS. In soybean it was estimated approximately 46,430 protein-coding genes spread over 20 chromosomes [3]. Other studies identified 205,614 tag SNPs that can be further used in association studies and QTLs mapping program [4]. Moreover Li et al. [5] analyzed 25 new and 30 previously published soybeans genomes from China accessions and identified a

total of 5,102,244 SNPs. Zhou et al [6] found associations for 10 selected regions and 13 previously uncharacterized agronomic loci. These studies highlighted the great power of the high-throughput technologies for GWAS.

A large number of GWAS are already available for several plant species, such barley [7, 8] rice [9] sorghum [10] pepper [5] and maize [11]. Aranzana et al. [12] identified in *Arabidopsis thaliana* previously known flowering time and pathogen resistance genes (*Rpm1*, *Rps5* and *Rps2*) in 95 accessions which have genome-wide polymorphism data available. Mamidi et al. [13] mapped 15 genes involved in iron metabolism in two populations of soybean.

In this study, using data from 27 Brazilian cultivars and 26 soybean lines from different origins, we performed a GWAS to identify important allelic and structural variations related to the resistance to three important soybean diseases: soybean cyst nematode (SCN) (*Heterodera glycines*) soybean root-knot nematode (RKN) (*Meloidogyne incognita*) and soybean stem canker (SSC) (*Diaporthe phaseolorum f. sp. meridionalis*).

Results and Discussion

Sequencing and variant calling

The genome sequence of 53 soybean accessions were used in this study being 27 from Brazil, six from United States and 20 from Asia (14 from China, 4 from Korea, and 2 from Japan). These materials are important soybeans used in breeding programs due the presence of a great number of resistance mechanism against the three diseases analyzed in this study (**Supplementary Table 1**).

The sequencing effort of these accessions generated around 9.7 billion of paired-end reads with 100 bp read length and an average of coverage of 13.86x the soybean genome. The average percentage of mapped reads on soybean genome for each accession was 91.76% which demonstrate that the resequencing was able to cover most part of soybean genome (**Supplementary Table 2**). We identified 4,511,750 SNPs spreaded in all chromosomes compared to the reference genome. Chromosome 15 and 18 had the highest number of SNPs and highest variant ratio per chromosome length with an average of a SNP every 145 and 137 bases, respectively (**Supplementary Table 3)**. The transition/transversion ratio (ts/tv ratio) was 1.85, with transitions being the most common nucleotide base change (**Supplementary Figure 1a**). Most of the SNPs were identified in intergenic region

(43.90%). In coding regions we found 90,117 SNPs in UTR regions 230,710 in introns and 175,141 in exons (**Supplementary Figure 1b**). The non-synonymous-to-synonymous ratio observed was 1.50. This value was lower than observed in other soybean study [4], however, higher than observed in sorghum [14] and rice [15]. A summary of the number of SNPs identified related to all diseases analyzed in this study can be visualized in **Table 1**. Moreover, the type and region of the modifications caused by SNPs for each biotic stresses of this study are showed in **Table 2**.

A large numbers of QTLs influence soybean cyst nematode race 1 resistance.

Soybean cyst nematode, caused by *Heterodera glycines* Ichinohe, is one of the most worldwide economically important soybean pests. In 2009, It was estimated losses of 120,048,000 bushels in United States [2]. The use of resistant cultivars associated with nonhost crop rotation is the best approach to decrease the impact of the pest. However, due the large number of races-type of SCN, the identification and use of resistance genes are being a great challenge for breeding programs around the world. In this study, we focus to find important SNPs related to resistance mechanism against SCN race 1 and 3.

For SCN race 1, we identified 1,462 SNPs related to the resistance against race 1 of SCN on chromosomes 1, 8, 9, 17, 19 and 20 (**Figure 1a**) with a significant correlation between phenotype and genotype data according to QQ-plot analysis (**Supplementary Figure 2).** The largest number of SNPs with meaningful markers correlation (r^2) and *p-values* were identified on chromosome 17 **(Table 3)**, which could mean an important mechanism to the resistance against SCN race 1 for these soybeans.

There are two important major genes mapped for SCN resistance: *Rhg1* and *Rhg4*. The *Rhg1* gene was mapped on chromosome 18, whereas *Rhg4* was mapped on chromosome 8 [16, 17]. In SCN major genes region, we just identified SNPs related to SCN race 1 resistance inside *Rhg4* gene. For chromosome 8, we found 29 SNPs, being eight inside *Rhg4* interval and the remaining in two QTLs regions described in other studies, one between 7.55 – 7.61 megabase (Mbp) [18, 19] and the other between 18.85 – 18.88 Mbp [20]. Susceptible cultivar BRS/GO 8660 had allelic variation similar to resistant materials for *Rhg4* region. Although BRS/GO 8660

is susceptible for SCN race 1, this cultivar had resistance against SCN race 3, which could explain the presence of this allele in *Rhg4* region. At the same chromosome, a similar pattern of *Rhg4* was observed in QTL region between 7.55 - 7.61 Mbp and other studies reported SCN race 1 QTLs inside this region [18, 19]. For the other QTL region, cultivar BRS/GO Chapadões did not present the resistance allele, which could mean the absence of this QTL in this cultivar. Moreover, SCN race 1 susceptible cultivars BRSMT Uirapuru and FT Cristalina presented the resistance allelic pattern, which could be the existence of this QTL for these cultivars, but the presence of only this resistance sequence is not enough to create a defense mechanism against SCN race 1.

On chromosome 1, 345 SNPs were found in inside two QTL regions. The first QTL was located between 0.41 - 0.96 Mbp. A large number of blocks with strong linkage disequilibrium (LD) in resistant materials can be observed in this chromosome (**Supplementary Figure 3**). Accessions BRSMT Pintado, P98Y11 and HN018 do not have the resistance allele for SCN race 1, which could mean the absence of this QTL in these three soybeans lines. For the other QTL region, located between 5.57 – 5.58 Mbp, we observed absence of resistance allele to BRSMT Pintado, Forrest, P98Y11, and HN018. Our findings suggested that these lines do not have important QTLs against SCN race 1 on chromosome 1, except for Forrest, where allelic variation is absent only in the second QTL. Yue et al [21] found QTLs associated with SCN race 5 inside both regions identified in our study. However, there is no QTL related to SCN race 1, which could mean a new QTL related with SCN race 1 resistance or the same QTL previously descripted studying for more than one SCN race.

Furthermore, we found 668 SNPs in two regions of the chromosome 17. The first QTL region was identified between 13.81 – 23.12 Mbp. Other study identified a QTL between SSR markers Satt574 and Satt543[22], an interval inside the first QTL found in this chromosome. Due the fact these QTLs were identified by genetic mapping, this could not be their real physical position. Thus, our results can be the same QTLs identified in other studies but in a physical position or a new major QTL identified for SCN race 1. As we described for chromosome 1, a large number of blocks with strong LD in resistant accessions were observed in this QTL region (**Supplementary Figure 3**). Additionally, the other QTL identified in this chromosome

was located between 34.75 - 34.91 Mbp. There are no previously QTL described for this region, which can be a new QTL related to resistance against SCN race 1.

Finally, we found 676 SNPs in an interval between 35.02 – 35.40 Mb on chromosome 20. Previously study identified a QTL inside the same region against SCN race 3 [23]. Thus, the QTL identified in our study can be a new QTL related to SCN race 1 resistance or the same QTL controlling the resistance for more than one race of SCN. Two Asian accessions, HN010 and HN011, had a susceptible allele in most of the SNPs inside this QTL region. Therefore, these accessions do not have allelic variations for this QTL against SCN.

***Rhg1* and *Rhg4* have great impact to soybean cyst nematode race 3 resistance**

For SCN race 3, we detected 2,153 SNPs on chromosomes 7, 8, 10, 11, 17 and 18 (**Figure 1b**). We identified significant correlation between phenotype and SNP data according to QQ-plot analysis (**Supplementary Figure 2b**). The largest numbers of SNPs were identified on chromosome 10. However, the most meaningful SNPs *p-values* and r^2 were identified on chromosome 18 (**Table 3**). In this chromosome, there are 261 SNPs in six putative regions with SNPs that could be related to the genetic resistance against SCN race 3. A total of 73 SNPs were identified close or inside *Rhg1* region, the highest number found in this chromosome. *Rhg1* is the most important resistance gene against SCN. Our results suggested that most of the resistant material used in this study had important modifications caused by SNPs inside *Rhg1*. There is a large block with some important SNPs related to resistance against to SCN inside *Rhg1* region (**Supplementary Figure 4**). This finding is important to breeding programs to select and insert SCN region to new materials without *Rhg1* resistant gene. At the same chromosome, there are other five regions with meaningful SNPs related to resistance mechanism. There are several studies descripting a large number of SCN QTLs inside this region [20–27]. Thus, the SNPs detected for this chromosome has a higher correlation with genetic resistance against the race 3 of SCN.

Chromosome 8 is another important chromosome with a large number of SCN resistance regions. In this study, we identified 515 SNPs in two putative resistant regions. For the interval correspondent to *Rhg4*, we identified 502 SNPs, being most of them upstream 5 kbp of the genes identified inside *Rhg4* region. We

also identified a large block with strong LD for SNPs related to SCN race 3 in resistant materials inside this region (**Supplementary Figure 4**). An important fact observed was the cultivars Anta 82, VMAX RR and HN021 did not have resistance allelic variation, which could mean the absence of *Rhg4* resistance alleles for these accessions (**Supplementary Figure 5**). In contrast, PI 424608 has the resistance allelic variation, which could mean this gene is present in this accession but it is necessary the presence of other genes, such as *Rhg1* to increase the resistance against SCN race 3. There is other important region between 18.51 – 18.61 Mbp with 13 SNPs related with SCN race 3 resistance. This region had an absence of resistance allele in Anta 82, BRSMT Pintado, P98Y11 and VMAX RR, which could mean is not an important source of resistance allelic variation for these cultivars. In contrast, Conquista, BRS Valiosa, BRSMG 850G RR and IAC 8 have the resistant allele. This result could suggest this region is not primordial to increase the resistance against SCN race 3, although a QTL for SCN resistance have been already identified on this region [20].

Several other meaningful SNPs were identified in other soybean chromosomes. On chromosome 7, we found 16 SNPs very close to the QTL related to SCN resistance descripted by Webb et al [28]. Due the fact they studied with a genetic map, our result could explain better the position of this QTL. A similar result can be observed on chromosome 10. We identified 948 SNPs in two intervals between 41.40 – 43.40 Mbp. Vuong et al [20] found a QTL for this region, which could represent the presence of important SNPs related with the resistance against SCN race 3. For chromosome 17, we detected 163 SNPs related with SCN resistance in an interval of 0.33 – 0.67 Mbp. Yue et al [21] identified a SCN QTL in the beginning of chromosome 17. Due the fact they used a genetic map, it is possible the QTLs identified in both studies are the same. Anta 82 does not have the resistance allelic pattern, which could mean this QTL is absent in this cultivar. Moreover, BRS 284, BRSMT Uirapuru and PI 567387 has the resistance allelic pattern, which could mean they have this resistant region, but the presence of this region is not enough to provide resistance against SCN race 3.

Finally, chromosome 11 had 246 SNPs related to resistance against SCN race 3 in the intervals of 30.65 - 30.67 and 32.00 - 32.87 Mbp. The resistance allele is present in most of the resistant Brazilian cultivars and the two American cultivars. In contrast, some Asian accessions have a mix of resistance/susceptible allele, which

could mean these QTLs are important for resistance to SCN race 3 in the present commercial cultivars. Ferdous et al. [25] and Yue et al [21] identified important QTLs related with SCN race 3 resistance on this region.

Major root-knot nematode resistance QTL detected on chromosome 10

Southern root-knot nematode is other important agronomical nematode for soybean due the large impact caused in soybean production. One of the most common and important specie is *Meloidogyne incognita*. The use of resistant cultivars associated with nonhost crop rotation is the main strategy to decrease the losses caused by RKN. Thus, the identification of QTLs related to the resistance of this pest is essential for breeding programs.

For this analysis, we identified 4,461 SNPs related with the resistance against RKN. All of the SNPs were detected on chromosome 10 in the interval between 0-2.2 Mbp (**Figure 1c**), with a strong correlation between phenotype and SNP data according to QQ-plot analysis (**Supplementary Figure 2c**). Most of them are upstream 5 kbp each genes identified in this QTL, with 2,314 SNPs. This region was previously described in other study with the presence of a major QTL resistance against RKN [29, 30]. There are two regions with the highest *p-values* and r^2 values (**Table 3**). For the first region, we observed that HN012, HN021 and HN022 do not present a resistant allelic pattern, which may mean that other region inside this QTL are controlling the resistance mechanism against RKN in these three accessions. The other region had a similar allelic variation pattern for all the resistant material. However, seven susceptible materials, Anta 82, HN026, HN003, HN004, HN005, HN018, and HN020 do not have a susceptible allele. This finding could mean that this region were not enough to generate the resistance mechanism for these accessions. We also identified a large number of LD blocks spreaded in most part of this QTL region with several SNPs related to RKN in resistant accessions (**Figure 2**). This finding can have a great impact in breeding programs for RKN resistance, since these SNPs can be useful for MAS.

***Rdm?* is one of the most important resistance gene against to soybean stem canker**

Soybean stem canker (SSC), caused by *Diaporthe phaseolorum* f. sp. *meridionalis* is a historical soybean important disease, being responsible for

meaningful losses in soybean crops. In 1994, soybean stem canker was responsible for losses of 1,800,000 metric tons in Brazil, being the major problem for Brazilian soybean crops at that times [31]. Actually, this disease is controlled by the introgression of resistance genes in elite cultivars and is present in most of cultivars released over the last 20 years. However, there is a low number of SNPs available related with SSC resistance and the identification of new SNPs will be very important for soybean breeding programs to confirm the gene introgression in new soybean cultivars.

In this study, we identified 2,014 SNPs related to the resistance against SSC. All SNPs were detected on chromosome 14 (**Figure 1d**), most of them in an interval between 1.54 – 2.03 Mbp, where *Rdm?* gene were mapped [32]. There is a strong correlation between phenotype and genotype data according to QQ-plot analysis (**Supplementary Figure 2d**). Moreover, the highest number of SNPs was identified upstream 5 kb of each gene detected inside this interval. There are 235 SNPs with the highest *p-value* and r^2 values **(Table 3)**. For these SNPs, we found an allelic pattern that differentiates all accessions according to the phenotypic response (**Figure 3**). All the susceptible accessions have an allelic variation compared with the reference genome, cv Williams 82. This result suggests that cv. Williams 82 also have the *Rdm?* resistance gene in its genome. Moreover, this region has a large number of SNPs with strong LD, between some blocks inside *Rdm?* region (**Supplementary Figure 6**). No SNPs were identified on chromosome 02 that carry *Rdm1*, *Rdm2* and *Rdm4* genes, as well as *Rdm3* gene on chromosome 14. The main reason about it may be the SSC race-type inoculum used in the phenotype evaluation. In this study, all the accessions were separated according to their response against the isolate CH8 of *Diaporthe phaseolorum var. meridionalis*. Thus, some genes should have a race specific response. For a better detection, it will be necessary new evaluations with others SSC race-type.

Functional annotation analysis involving resistant genes reveals important modifications in soybean

By identifying SNPs and modified sequences within the genome of the accessions used in this study, it becomes possible to analyze the influence of these nucleotide mutations. Therefore, according to gene information in databases

obtained on SoyKB website, an enrichment analysis of these modified genes was made.

A large number of genes have important allelic variations detected in this study. For the major resistance regions, we identified SNPs inside *Rhg1*, *Rhg4* and *Rdm?* genes loci. For *Rdm?* region, we identified 73 non-synonymous modifications in exons caused by SNPs in susceptible lines for 25 genes. The non-synonymous SNPs with the best *p-values* were observed in two serine-threonine protein kinases (*Glyma.14g026300* and *Glyma.14g026700*), a leucine-rich repeat receptor-like protein kinase related to protein phosphorylation (Glyma.14g026500), a PH domain leucine-rich repeat-containing protein phosphatase 1 (*Glyma.14g024400*), a RNA helicase (*Glyma.14g024300*), a methyltransferases (*Glyma.14g026600*) and a purple acid phosphatase (*Glyma.14g024700*) (**Figure 3**). Plant-receptor-like serine/threonine kinase was one of the first genes cloned and associated to defense mechanisms, which plays a key role in signal transduction pathway in plants [33, 34]. Thus, these non-synonymous mutations in coding regions of the serine-threonine protein kinases identified in this study may lead to candidate genes for *Rdm?*. Other important non-synonymous SNPs on SSC resistance region were observed in a heat shock protein 70KDA (*Glyma.14g024200*), cytochrome P450 (*Glyma.14g027600*), putative translation initiation inhibitor UK114/IBM1 (*Glyma.14g023600*) and a DNA Mismatch repair ATPase MSH5 (*Glyma.14g022400*). In susceptible accessions, we also identified genes with SNPs responsible for splice site regions modifications of two genes: *Glyma.14g026600* and *Glyma.14G027100*. Furthermore, there were six SNPs in 5' UTR regions that may generate a new start codon in three genes: a gene related with rab GTPase activator activity (*Glyma.14g022300*), a heat shock protein 70KDA (*Glyma.14g024200*) and a PH domain leucine-rich repeat-containing protein phosphatase 1 (*Glyma.14g024400*). Finally, the cytochrome P450 had an allelic variation responsible for the presence of a new stop codon in coding region. In this case, most of the resistant material had this modification that may be related with SSC resistance and may be this sudden gene function change an important mechanism against SSC.

We also found 43 SNPs inside *Rhg1* interval, associated to resistance against SCN race 3. *Rhg1* is the most important resistance gene against SCN. Our results suggested that most of the resistant material used in this study had a similar SNP pattern on this region, with several modified genes. There were nine genes with

sequence modifications due the presence of a SNP inside *Rhg1* region, but only four had gene ontology description. Inside *Rhg1* region there were five genes with SNPs in upstream, downstream or coding gene regions (**Supplementary Figure 5**). *Glyma.18g022400* (encoding a predicted amino acid transporter), *Glyma.18g022500* (α-soluble N-ethylmaleimide-sensitive factor attachment protein - α-SNAP) and *Glyma.18G022600*, were identified and related with the resistance against SCN race 3 in other studies [35, 36]. Moreover, Cook et al [35] described the enhancing of the SCN resistance due modifications of the gene *Glyma.18g022500*. A non-synonymous mutation in coding region of resistant lines, that changes aspartic acid to tyrosine, was also found in this gene in our study, similar to the results obtained by Cook et al [36]. For *Glyma.18g022400* and *Glyma.18G022600*, we only detected SNPs downstream of the gene position. This finding associated with other studies reinforces the importance of the SNPs identified for SCN resistance in this study.

Finally, *Rhg4* is other important SCN major resistance gene identified on chromosome 8. We found 404 SNPs in *Rhg4* region related to resistance against SCN race 1 and 3. For race 3, we identified 398 SNPs related to 33 genes. There is a mutation observed on splice donor site region for *Glyma.08g107900* (a predicted histone H3) in most of the resistant materials. We also identified important SNPs that caused non-synonymous sequence modifications inside coding region of *Glyma.08g106500* (a gene with a PPR repeat domain) and *Glyma.08g107700* (a leucine-rich repeat receptor-like protein kinase). Kandoth et al. [37] described leucine-rich repeat receptor-like and PR proteins as an important defense mechanism in *Rhg1* against SCN. Puthoff et al [38] showed a transcription response to SCN for several genes, including PPR protein. Itahl et al. [39] showed that leucine-rich repeat receptor-like kinases are both up and down-regulated under SCN infection. However, Liu et al. [40] concluded that the leucine-rich repeat receptor-like kinase found in *Rhg4* is not a gene for SCN resistance. According to Liu et al. [41], a serine hydroxymethyltransferase is the major responsible for the *Rhg4* reaction against SCN. We identified a non-synonymous SNP in coding region inside this gene (*Glyma.08g108900*), which should be related to *Rhg4* reaction to SCN. Moreover, we found more non-synonymous mutations in other genes in this region, such glucosyl/glucuronosyl transferases genes (*Glyma.08g107500* and *Glyma.08g107600*), a peptidyl-prolyl cis-trans isomerase with a tetratricopeptide

repeat domain (*Glyma.08g106700*), UDP-glucosyl transferase (*Glyma.08g107500*) and a kinesin-like protein (*Glyma.08g106400*).

Additionally to major resistance genes information, there are a large number of genes inside QTLs with meaningful SNPs related to SCN resistance (**Supplementary Table 4**). Several studies described the importance of some of the genes that were identified with non-synonymous mutations in our study. Ithal et al [39] found a differential expression in BZIP/Myb domain genes, pathogenesis-related proteins, leucine-rich repeat proteins, transcription factors and cytochrome P450. Moreover, Vaghchhipawala et al [27] showed a enhancing on expression in cyclins, heat shock proteins, and pathogenesis-related proteins. Other studies showed that a large number of stress and defense-related genes were related with SCN defense, such as leucine-rich repeat, heat shock protein, pathogenesis-related proteins and Myb domains. Such results reinforce the importance of non-synonymous SNPs inside these genes of the resistant materials.

Finally, important sequence modifications were observed in RKN analysis. A total of 3,221 SNPs were identified related to 152 important genes within the major RKN QTL detected in this study. Xu et al [29] described *Glyma10g02150* and *Glyma10g02160*, a pectin methylesterase inhibitor and a pectin methylesterase inhibitor -pectin methylesterase, as the most important genes involved with RKN resistance. The first gene was removed in the new version of soybean genome and the second becomes *Glyma.10G017200*. For the second gene, we just found synonymous mutations in coding region and modifications in 3' UTR. However, we identified another gene close to both genes with a similar function: *Glyma.10G017100*. For this specific gene, we found important sequence modifications caused by SNPs (**Supplementary Figure 7**). We found five non-synonymous modifications on coding sequences of resistant materials. In addition, we identified upstream of these five non-synonymous SNP, a mutation that create a stop codon on resistant accessions. This finding may mean that post-translational modifications due the existence of a new stop codon in a coding region could be related with partial resistance against RKN.

Several genes related to resistant mechanism against biotic factors were identified with non-synonymous mutation in coding sequences of resistant materials against RKN (**Supplementary Table 5**). In Arabidopsis, Fuller et al. [42] described the differential expression of glycosyl transferases and genes with zinc finger

domain. Moreover, the same study related that genes with Myb domain could be repressed during root-knot nematode infection. Potenza et al. [43] showed that NADH-dependent oxidoreductases genes might play a role for RKN resistance, by scavenging reactive oxygen species originated of the pathogen interaction with the plant. Moreover, Bakhetia et al [44] demonstrate a decrease of number of eggs due the silencing of oxidoreductases. Other studies illustrated the importance of heat shock proteins in RKN infection. Escobar et al [45] showed that heat shock elements are related to the activation of giant cells and heat shock transcription factors may mediate the response against RKN in tobacco. Moreover, Lopes-Caitar et al. [46] identified heat shock proteins family related to javanese RKN infection. In our study, we identified non-synonymous mutations in coding regions and new start codons in 5' UTR region inside a DNAJ heat shock n-terminal domain-containing protein and a heat shock transcription factor of resistant accessions. Finally, Ibrahim et al [47] showed the importance of cyclin in soybean, due the increase of its action after 12 days after RKN infection. The large number of non-synonymous SNPs inside known nematode resistant genes demonstrates the importance of this QTL region on RKN resistance.

Additionally, we identified important sequences modifications in splice site, start codons caused by SNPs related to RKN resistance. In splice site, we identified modifications on donor and acceptor sites for iron/ascorbate family oxidoreductases (*Glyma.10G007500*) and a glycosyl hydrolases family 28 (*Glyma.10G016900*), respectively. Furthermore, we found a SNP mutation that generates a possible pseudogene due the loss of a start codon in resistant material in a late embryogenesis abundant-related protein (*Glyma.10G014200*). Finally, we identified modifications in 5' UTR region that may create a new start codon nine genes: *Glyma.10G001200*, *Glyma.10G001400*, *Glyma.10G003100*, *Glyma.10G003500*, *Glyma.10G003600*, *Glyma.10G007800*, *Glyma.10G009500*, *Glyma.10G012400* and *Glyma.10G014000*. The presence of new start codons associated with non-synonymous mutations in resistant materials may have a key role for RKN defense mechanism in this QTL.

CNV as an important tool to detect resistant mechanism

CNVs are structural modifications that caused variation in copy-number of sequences fragments in specific genome region. Several studies were described with

CNVs analysis in a large number of plants, such as *Arabidopsis thaliana* (DEBOLT et al, 2010), barley (MUNOZ-AMATRIAIN et al., 2013), maize (SPRINGER et al., 2009) and soybean (Zhou et al., 2015). Thus, CNVs studies are extremely important and may have a key role for plant diseases defense mechanism against a specific pathogen. In this study, we analyzed 27 Brazilian lines and some resistant materials to find important CNVs that could be related to resistance mechanism against SCN, RKN and SSC.

Our preliminary results detected several CNVs inside important resistance regions on chromosome 08, 10, 11, 17 and 18. The accession Forrest had the most number of exclusive CNVs in SCN resistance regions (**Supplementary Figure 8**). As one of the most important source of resistance against this nematode, such CNVs identified only in Forrest may be related to genes having a great importance to defense mechanism against SCN.

Moreover, we also found some CNVs inside *Rhg1* and *Rhg4* genomic regions present only in resistant accessions. For *Rhg4* interval, we identified three important CNVs (**Figure 4**). Two of these were deletions observed in BRS/GO Chapadões, Forrest, HN002, HN003, HN004, HN005, HN011, HN015, HN018, and HN021. Additionally, we identified insertions in the same interval in BRS/GO 8660, BRSMT Pintado, P98Y11, G93-9223, HN005, HN004, HN010, and HN022. In contrast, in *Rhg1* interval, we found an insertion in Anta 82, G93-9223, VMAX RR, and HN021 (**Figure 4**). Cook et al. [35, 36] confirm an increase of *Rhg1* resistance due the presence of CNVs in *Rhg1* region. Thus, such structural modifications in *Rhg1* and *Rhg4* regions may have a great importance in increase of resistance in such materials. Moreover, our results in *Rhg4* regions, associated with GWA analysis could suggest a non-influence or absence of *Rhg4* gene in resistant cultivars Anta 82 and VMAX RR.

In other regions containing QTLs associated to SCN resistance, we also detected CNV modifications (**Supplementary Figure 9)**. Divergent CNV pattern between resistant and susceptible materials can also be observed at the end of the chromosome 10. We identified deletions for resistant materials Anta 82, BRS/GO Chapadões, BRSMT Pintado, Forrest, HN002, HN003, HN004, HN005, HN008, HN010, HN011, HN015, HN018, HN021, and HN022. In contrast, cultivars BRS Valiosa, BRS/GO 8360, BRSMG 850G RR, Conquista, IAC 8 and Santa Rosa, susceptible lines against SCN, showed insertions inside the same region. This

finding could be directly related to SCN resistance mechanism. Another interesting pattern was observed on chromosome 11. We found deletions and insertions in 16 Brazilian susceptible cultivars. From the abroad lines, only HN005 and HN018 had similar pattern. For this region, according to our GWA analysis results, these two accessions do not have resistant allele, which could suggest that modifications in this region could be responsible for an increase of susceptibility in soybean accessions. Similar pattern can be observed in the beginning of the chromosome 18, which we found an deletion in ten susceptible Brazilian lines that could be related to an increase of SCN susceptibility in soybeans. Finally, on chromosome 17, there are more CNVs regions that could be related to SCN defense mechanisms. In the beginning of the chromosome, we found two CNVs in nine Brazilian susceptible soybeans. Moreover, we identified CNVs modifications inside the interval that we detected SNPs related to a putative resistance QTL against SCN race 1. For resistant materials, we found deletions in eight regions. In contrast, we also identified other eight insertions in Brazilian susceptible lines. This divergent pattern observed between these materials may show an important control resistance region against SCN.

We also identified important CNVs in resistance regions against RKN. We investigated regions on chromosomes 8, 10, 13 and 18, due the existence of known RKN QTLs described in other studies. For these regions that we identified important SNPs related to the resistance QTL against this disease, we identified a large numbers of deletions for cultivar Forrest, most of them in non-coding regions. However, two glycosyl hydrolase (*Glyma.10G016900* and *Glyma.10G017000*) could be affected due a deletion upstream the gene (**Figure 4**). Such deletions, associated to the fact that cv. Forrest is one of the resistant materials, could infer that an alternative action of these genes could be related to RKN resistance mechanism. In other QTL regions on chromosome 8, 10, 13 and 18, we identified more exclusive deletions in cv. Forrest, which may represent the absence of important regions could increase the resistance against RKN. Moreover, we identified an important deletion on the 0.262 – 0.267 Mbp region on chromosome 10, shared with BRS Valiosa RR, BRSMG 850G RR, CD 201 and Conquista, and located upstream the transcription initiation factor *Glyma.10g002500* (**Figure 4**). The presence of this deletion could be related with the resistance of these lines against RKN.

Finally, we found CNVs in resistant regions against SSC. Chromosome 2 and 14 were investigated due the presence of *Rdm* genes: *Rdm1*/*Rdm2*/*Rdm4* on chromosome 2 and *Rdm3*/*Rdm?* on chromosome 14. According to our analysis, accession HN020 has exclusive CNVs modifications on *Rdm3* and *Rdm?* regions (**Supplementary Figure 10**). For the putative *Rdm?* region on chromosome 14, we identified two deletions in HN020 line (**Figure 4**). One of these can be affecting sequences of two genes: a heat shock protein 70KDA (*Glyma.14g024200*) and a RNA Helicase (Glyma.14g024300). However, this deletion does not appears to influence HN020 resistance to SSC, due the fact that sequences modifications were not observed previously in resistant accessions when compared with reference genome, cv. Williams 82, probably a resistant cultivar. Moreover, we found more exclusives deletions inside *Rdm3* intervals to the same accession. We detected another two CNVs modifications inside *Rdm?* and *Rdm3* regions. The first was an insertion between 1.68 - 1.70 Mbp observed in BRS 284, BRSMT Uirapuru and CD 201. This could be responsible for an increase of resistance in such lines. Another insertion was identified between 3.36-3.38 Mbp in accessions BRS Valiosa, BRS/GO Chapadões, BRSMG 850G RR, IAC 8, Conquista and HN020. The cultivar BRS/GO Chapadões is the only susceptible material to SSC with this modification. However, this line has a moderate field resistance to SSC (Carlos Alberto Arrabal Arias, personal communication). With the lack of information about some accessions pedigree and the physical position of *Rdm3*, this result could suggest the presence of a region related to resistance mechanism against SSC and may be an important region for BRS/GO Chapadões field resistance against SSC.

On chromosome 2 we also found more CNVs inside putative *Rdm* regions of resistant materials (**Supplementary Figure 11**). Two regions inside *Rdm4* interval had deletions for Anta 82, BRS 360 RR, BRS Valiosa RR, NA 5909 RG, P98Y11 and VMAX RR. Moreover, other two regions inside *Rdm1* had CNVs modifications for some resistant accessions. For one region, we identified deletions inside the accessions BRS 284, BRS Valiosa, BRS/GO 8660, BRSMG 850G RR and NA 5909 RG and insertions for Anta 82, BRS/GO 8360, Doko and HN003. For another region, we detected insertions for Anta 82, BRS/GO 8360 and Doko and deletions for BRS 284, IAC 8, Conquista and NA 5909 RG. Similar to the information discussed above, the lack of the physical position of *Rdm1* and *Rdm4* difficult to infer if these CNVs regions are inside *Rdm* genes. However, the presence of a pattern in a large number

of resistant materials can suggest a relative importance of these regions to SSC resistance. Finally, there was an insertion observed in Doko and deletions in accessions BRS 232, BRS 360 RR, BRS Valiosa RR, BRSMG 850G RR, Conquista, HN003, HN020 and HN021 between 0.25 – 0.26 Mbp. This region is not in a gene region, but the presence of CNVs inside this region in resistant materials can represent the presence of an important mechanism for plant resistance against SSC.

Conclusion

This study provides promising applications on breeding programs for disease resistance. The discovery of important SNPs related to a large number of genomic regions associated with the resistance against RKN, SSC, SCN race 1 and 3 can have a direct impact in marker assisted programs for disease resistance.

Our data suggest a predominant role of allelic variations in *Rdm?* locus, being an important source of resistance against soybean stem canker caused by the isolate CH8 of Diaporthe phaseolorum f. sp. meridionalis. Similarly, the resistance against SCN was strongly associated to allelic variations in *Rhg1*, *Rhg4* and some QTLs previously described in other studies. Finally, the allelic variations observed in the QTL of chromosome 10 previously described in other works were the major resistance mechanism against RKN. Furthermore, the CNVs identified in this study suggested specific patterns related to resistance against SCN, as well as important modifications in loci associated to SSC resistance mechanisms.

Our study opens a new horizon for genotyping and the identification of accessions with resistance against the three studied diseases. However, a validation process for the SNPs will be necessary to confirm the results obtained in this study. In addition, a depth study of the detected CNVs may be made due the possibility of large influence of these structural modifications in defense mechanism against these soybean pathogens.

Conflict of Interests

The authors declare that they have no conflict of interests

Acknowledgements

We greatly appreciate financial support from the Coordination for the Improvement of Higher Level or Education program (CAPES). We thank the

plant biotechnology and bioinformatics laboratory members at Embrapa Soja in Brazil for supporting this study. Furthermore, we thank the Molecular Genetics & Soybean Genomics Laboratory (Division of Plant Sciences) and Digital Biology Laboratory (Computer Sciences Department) at the University of Missouri in the United States for supporting the doctoral student exchange program and this research.

Materials and Methods

Plant material and sequencing

Twenty-seven Brazilian cultivars were selected for this study and data were used for all the analysis, whereas the remaining accessions from United States and Asia were selected according to phenotype information available. Young leaf tissue sample of each 27 Brazilian cultivars were collected during stage V3. The genomic DNA was isolated for each sample with the Qiagen Mini Plant DNeasy kit (Qiagen Inc., Valencia, CA, USA), following the manufacturer's instructions. The quality of the extracted DNA was checked through a 0.8% agarose gel stained with etidium bromite and the quantification was measure through the Qubit 2.0 (Life Technologies, Invitrogen division, Darmstadt, Germany) fluorometer. The Brazilian cultivars samples were sent to FASTERIS Company, Switzerland, for sequencing. The remained soybeans genomics were kindly provided by the Molecular *Genetics* and *Soybean* Genomics Laboratory from the University of Missouri and were sequenced through the company BGI on China. The sequencing efforts were done on the Illumina Hiseq 2000 platform generating 100 bp reads paired-end with an expected coverage of 15x the soybean genome.

Phenotype information

The Phenotypic data information for disease resistance was provided by Embrapa Soja and Molecular Genetics and Soybean Genomics Laboratory from the University of Missouri. SSC phenotypic resistance information was provided by plant-pathology laboratory from Embrapa Soja for all accessions. The accessions were separated according to their response against SSC in resistant, moderately resistant and susceptible. In additional, the nematology laboratory from Embrapa Soja provided information about the resistance of Brazilian accessions under SCN and RKN infection. Phenotypic resistance information of American and Asian accessions against SCN and RKN were kindly provided by

the Molecular *Genetics* and *Soybean* Genomics Laboratory from the University of Missouri. SCN phenotype information separated accessions into resistance, moderately resistant, moderately susceptible and susceptible, whereas RKN phenotype information clustered materials into resistance or susceptible.

SNPs and indels detection

The resequenced soybean accessions were mapped with the new version of the soybean reference genome (Gmax_275_Wm82.a2.v1) through the alignment program Burrows-Wheeler Aligner (BWA) [48]. After mapping the aligned reads were processed through Piccard tools version 1.107 to remove duplicate values and a binary file of extension bam representing the assembled genome of each resequenced species were generated. For SNPs/indels calling we used the Genome Analysis Toolkit (GATK) version 3.0 [49]. This toolkit was responsible to make a local realignment in indels region and a qualitative recalibration for the purpose to generate a bam file with fewer errors for each sample. Thus the new bam files generated were used to SNPs/indels calling of the genome. In both cases we used the HaplotypeCaller module of the GATK.

The analysis was conducted using bioinformatics NGS resequencing data analysis workflow [50] developed in SoyKB for SNP calling and was conducted using XSEDE as the computing infrastructure, iPlant as the data and cloud infrastructure [51], and the Pegasus workflow systems [52] to control and coordinate the data management and computational tasks.

Genome-wide association analysis

For genome-wide association analysis we used GAPIT program (LIPKA et al. 2012). A Mixed Linear Model was selected for this analysis with EMMA as a kinship matrix algorithm. Moreover, we used a major allele imputation method for missing data about, PCA total of 3 and a minimum p-value of 7.19-e10^{-4} to detect significant SNPs.

Linkage disequilibrium detection

To measure the linkage disequilibrium (LD) level of the soybeans accessions the correlation coefficient (r^2) of the alleles was calculated using Haploview (Barrett et al. 2005). The parameters of the program were set as follow: −maxdistance 1000

−dprime –memory 2000 −minMAF 0.1 −hwcutoff 0.001. To identify the LD block we included the parameters '−blockoutput GAB –pairwiseTagging' were included to the program.

Annotated information, functional classification and prediction effect of gene with meaningful SNPs.

The functional classification of the genes where allelic variations had been detected was based on the snpEff program [53]. An enrichment analysis of these modified genes detected through snpEff were made through the website agriGO [54] and SoyKB [55].

Copy-Number Variation (CNV) identification

We used Copy Number estimation by a Mixture Of Poissons (cn.MOPS) version 1.10.0 [56] for CNV detection.

References

1. Lee G-A, Crawford GW, Liu L, Sasaki Y, Chen X: **Archaeological soybean (Glycine max) in East Asia: does size matter?**. *PLoS One* 2011, **6**:e26720.
2. Koenning SR, Carolina N, Box PO, Wrather JA: **Suppression of Soybean Yield Potential in the Continental United States by Plant Diseases from 2006 to 2009 Plant Health Progress Plant Health Progress**. 2010(October):2006–2011.
3. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, et al.: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**:178–83.
4. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS-M, Zhang G: **Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.** *Nat Genet* 2010, **42**:1053–9.
5. Li Y, Zhao S, Ma J, Li D, Yan L, Li J, Qi X, Guo X, Zhang L, He W, Chang R, Liang Q, Guo Y, Ye C, Wang X, Tao Y, Guan R, Wang J, Liu Y, Jin L, Zhang X, Liu Z, Zhang L, Chen J, Wang K, Nielsen R, Li R, Chen P, Li W, Reif JC, et al.: **Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing.** *BMC Genomics* 2013, **14**:579.
6. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee S-H, Wang W, Tian Z: **Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean.** *Nat Biotechnol* 2015(April 2014).

7. Massman J, Cooper B, Horsley R, Neate S, Dill-Macky R, Chao S, Dong Y, Schwarz P, Muehlbauer GJ, Smith KP: **Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm**. *Mol Breed* 2010, **27**:439–454.

8. Pasam RK, Sharma R, Malosetti M, van Eeuwijk F a, Haseneyer G, Kilian B, Graner A: **Genome-wide association studies for agronomical traits in a world wide spring barley collection.** *BMC Plant Biol* 2012, **12**:16.

9. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q-F, Li J, Han B: **Genome-wide association studies of 14 agronomic traits in rice landraces.** *Nat Genet* 2010, **42**:961–7.

10. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S: **Population genomic and genome-wide association studies of agroclimatic traits in sorghum.** *Proc Natl Acad Sci U S A* 2013, **110**:453–8.

11. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas M a, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB: **Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population.** *Nat Genet* 2011, **43**:163–8.

12. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M: **Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes.** *PLoS Genet* 2005, **1**:e60.

13. Mamidi S, Chikara S, Goos RJ, Hyten DL, Annam D, Moghaddam SM, Lee RK, Cregan PB, McClean PE: **Genome-Wide Association Analysis Identifies Candidate Genes Associated with Iron Deficiency Chlorosis in Soybean**. *Plant Genome J* 2011, **4**:154.

14. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frère C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J: **Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum.** *Nat Commun* 2013, **4**:2320.

15. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W: **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nat Biotechnol* 2012, **30**:105–11.

16. Concibido VC, Lange D, Denny RL, Orf J, Young N: **Genome mapping on soybean cyst nematode resistance genes in "Peking", PI 90763, and PI 88788 using DNA markers.** *Crop Sci* 1997, **37**:258–264.

17. Concibido VC, Denny RL, Boutin SR, Hautea R, Orf JH, Young ND: **DNA marker analysis of loci underlying resistance to soybean cyst nematode (Heterodera glycines Ichinohe)**. *Crop Sci* 1994, **34**:240–246.

18. Guo B, Sleper D a., Nguyen HT, Arelli PR, Shannon JG: **Quantitative Trait Loci underlying Resistance to Three Soybean Cyst Nematode Populations in Soybean PI 404198A**. *Crop Sci* 2006, **46**:224.

19. Vuong TD, Sleper D a., Shannon JG, Wu X, Nguyen HT: **Confirmation of quantitative trait loci for resistance to multiple-HG types of soybean cyst nematode (Heterodera glycines Ichinohe)**. *Euphytica* 2011, **181**:101–113.

20. Vuong TD, Sleper D a, Shannon JG, Nguyen HT: **Novel quantitative trait loci for broad-based resistance to soybean cyst nematode (Heterodera glycines Ichinohe) in soybean PI 567516C.** *Theor Appl Genet* 2010, **121**:1253–66.

21. Yue P, Arelli R, Sleper A: **Molecular characterization of resistance to Heterodera glycines in soybean PI 438489B**. *Theor Appl Genet* 2001, **14**:921–928.

22. Kazi S, Shultz J, Afzal J, Hashmi R, Jasim M, Bond J, Arelli PR, Lightfoot D a: **Iso-lines and inbred-lines confirmed loci that underlie resistance from cultivar "Hartwig" to three soybean cyst nematode populations.** *Theor Appl Genet* 2010, **120**:633–44.

23. Winter SMJ, Shelp BJ, Anderson TR, Welacky TW, Rajcan I: **QTL associated with horizontal resistance to soybean cyst nematode in Glycine soja PI464925B.** *Theor Appl Genet* 2007, **114**:461–72.

24. Arriagada O, Mora F, Dellarossa JC, Ferreira MFS, Cervigni GDL, Schuster I: **Bayesian mapping of quantitative trait loci (QTL) controlling soybean cyst nematode resistant**. *Euphytica* 2012, **186**:907–917.

25. Ferdous SA, Watanabe S, Suzuki-Orihara C, Tanaka Y, Kamiya M, Yamanaka N, Harada K: **QTL Analysis of Resistance to Soybean Cyst Nematode Race 3 an Soybean Cultivar Toyomusume**. *Breed Sci* 2006, **56**:155–163.

26. Guo B, Sleper D a, Arelli PR, Shannon JG, Nguyen HT: **Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763.** *Theor Appl Genet* 2005, **111**:965–71.

27. Vaghchhipawala Z, Bassüner R, Clayton K, Lewers K, Shoemaker R, Mackenzie S: **Modulations in Gene Expression and Mapping of Genes Associated with Cyst Nematode Infection of Soybean**. *Am Phytopathol Soc* 2001, **14**:42–54.

28. Webb DM, Baltazar BM, Rao-Arelli AP, Schupp J, Clayton K, Keim P, Beavis WD: **Genetic mapping of soybean cyst nematode race-3 resistance loci in the soybean PI 437 . 654**. *Theor Appl Genet* 1995, **14**:574–581.

29. Xu X, Zeng L, Tao Y, Vuong T, Wan J, Boerma R, Noe J, Li Z, Finnerty S, Pathan SM, Shannon JG, Nguyen HT: **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* 2013, **110**:13469–74.

30. Li Z, Jakkula L, Hussey RS, Tamulonis JP, Boerma HR: **SSR mapping and confirmation of the QTL from PI96354 conditioning soybean resistance to southern root-knot nematode**. *Theor Appl Genet* 2001, **103**:1167–1173.

31. Wrather JA, Anderson TR, Arsyad DM, Gai J, Ploper LD, Porta-Puglia A, Ram HH, Yorinori JT: **Special Report Soybean Disease Loss Estimates for the Top 10 Soybean Producing Countries in 1994**. *Plant Dis* 1997, **81**:107–110.

32. Shearin ZP: **ENHANCING RESISTANCE TO SOUTHERN STEM CANKER AND SOUTHERN ROOT-KNOT NEMATODE IN SOYBEAN**. 2007.

33. Martin GB, Brommonschenkel SH, Chunwongse J, Frary a, Ganal MW, Spivey R, Wu T, Earle ED, Tanksley SD: **Map-based cloning of a protein kinase gene conferring disease resistance in tomato.** *Science (New York, N.Y.)* 1993:1432–1436.

34. Zhou J, Loh YT, Bressan R a., Martin GB: **The tomato gene Pti1 encodes a serine/threonine kinase that is phosphorylated by Pto and is involved in the hypersensitive response**. *Cell* 1995, **83**:925–935.

35. Cook DE, Bayless AM, Wang K, Guo X, Song Q, Jiang J, Bent AF: **Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode.** *Plant Physiol* 2014, **165**:630–647.

36. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless A, Wang J, Hughes TJ, K.Willis D, Clemente T, Diers BW, Jiang J, Hudson ME, Bent AF: **Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean**. *Science (80-)* 2012, **338**:1206–1209.

37. Kandoth PK, Ithal N, Recknor J, Maier T, Nettleton D, Baum TJ, Mitchum MG: **The Soybean Rhg1 locus for resistance to the soybean cyst nematode Heterodera glycines regulates the expression of a large number of stress- and defense-related genes in degenerating feeding cells.** *Plant Physiol* 2011, **155**(April):1960–1975.

38. Puthoff DP, Ehrenfried ML, Vinyard BT, Tucker ML: **GeneChip profiling of transcriptional responses to soybean cyst nematode, Heterodera glycines, colonization of soybean roots**. *J Exp Bot* 2007, **58**:3407–3418.

39. Ithal N, Recknor J, Nettleton D, Maier T, Baum TJ, Mitchum MG: **Developmental transcript profiling of cyst nematode feeding cells in soybean roots.** *Mol Plant Microbe Interact* 2007, **20**:510–525.

40. Liu X, Liu S, Jamai A, Bendahmane A, Lightfoot D a., Mitchum MG, Meksem K: **Soybean cyst nematode resistance in soybean is independent of the Rhg4 locus LRR-RLK gene**. *Functional and Integrative Genomics* 2011:539–549.

41. Liu S, Kandoth PK, Warren SD, Yeckel G, Heinz R, Alden J, Yang C, Jamai A, El-Mellouki T, Juvale PS, Hill J, Baum TJ, Cianzio S, Whitham S a., Korkin D, Mitchum MG, Meksem K: **A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens**. *Nature* 2012, **492**:256–260.

42. Fuller VL, Lilley CJ, Atkinson HJ, Urwin PE: **Differential gene expression in Arabidopsis following infection by plant-parasitic nematodes Meloidogyne incognita and Heterodera schachtii**. *Mol Plant Pathol* 2007, **8**:595–609.

43. Potenza CL, Thomas SH, Higgins E a, Sengupta-Gopalan C: **Early Root Response to Meloidogyne incognita in Resistant and Susceptible Alfalfa Cultivars.** *J Nematol* 1996, **28**:475–484.

44. Bakhetia M, Charlton W, Atkinson HJ, McPherson MJ: **RNA interference of dual oxidase in the plant nematode Meloidogyne incognita.** *Mol Plant Microbe Interact* 2005, **18**:1099–1106.

45. Escobar C, Barcala M, Portillo M, Almoguera C, Jordano J, Fenoll C: **Induction of the Hahsp17.7G4 promoter by root-knot nematodes: involvement of heat-shock elements in promoter activity in giant cells.** *Mol Plant Microbe Interact* 2003, **16**:1062–1068.

46. Lopes-Caitar VS, de Carvalho MCCG, Darben LM, Kuwahara MK, Nepomuceno AL, Dias WP, Abdelnoor R V, Marcelino-Guimarães FC: **Genome-wide analysis of the Hsp20 gene family in soybean: comprehensive sequence, genomic organization and expression profile analysis under abiotic and biotic stresses.** *BMC Genomics* 2013, **14**:577.

47. Ibrahim HMM, Hosseini P, Alkharouf NW, Hussein EH a, Gamal El-Din AEKY, Aly M a M, Matthews BF: **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* 2011, **12**:220.

48. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

49. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo MA: **The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data**. 2010:1297–1303.

50. Liu Y, Khan SM, Wang J, Chen S, Rynge M, Wang J, Santos JVM dos, Valliyodan B, Merchant N, Nguyen HT, Xu D, Joshi T: **Large Scale NGS resequencing data analysis workflow for soybean germplasm using iPlant, XSEDE and SoyKB framework**. *Bioinformatics* 2014, **in press**.

51. Goff S a, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, Muir A, Merchant N, Lowry S, Mock S, Helmke M, Kubach A, Narro M, Hopkins N, Micklos D, Hilgert U, Gonzales M, Jordan C, Skidmore E, Dooley R, Cazes J, McLay R, Lu Z, Pasternak S, Koesterke L, Piel WH, et al.: **The iPlant Collaborative: Cyberinfrastructure for Plant Biology.** *Front Plant Sci* 2011, **2**(July):34.

52. Deelman E, Singh G, Su M, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman GB, Good J, Laity A, Jacob JC, Katz DS: **Pegasus : A framework for mapping complex scientific workflows onto distributed systems**. *Sci Program* 2005, **13**(January):219–237.

53. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3**. *Landes Biosci* 2012, **6**:80–92.

54. **ANALYSIS TOOLKIT FOR THE AGRICULTURAL COMMUNITY (agriGO)** [http://bioinfo.cau.edu.cn/agriGO/analysis.php]

55. Joshi T, Patil K, Fitzpatrick MR, Franklin LD, Yao Q, Cook JR, Wang Z, Libault M, Brechenmacher L, Valliyodan B, Wu X, Cheng J, Stacey G, Nguyen HT, Xu D: **Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics.** *BMC Genomics* 2012, **13 Suppl 1**(Suppl 1):S15.

56. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S: **cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate.** *Nucleic Acids Res* 2012, **40**:e69.

Figure and Table Legends

Table 1. Total of SNPs related to important disease resistance regions.

Table 2. Summary of regions and type of modifications caused by SNPs for each biotic stresses

Table 3. SNP markers strongly associated with the four biotic stresses analysis in this study.

Figure 1. Manhattan plot from the three biotic stresses analyzed in this study. a SCN race 1 results b SCN race 3 results, c RKN results and d SSC results.

Figure 2. Linkage disequilibrium graphic for RKN QTL regions. The green lines represent the SNPs related to SCN resistance.

Figure 3. Non-synonymous SNPs strongly associated with SSC resistance in *Rdm?* region of commercial accessions.

Figure 4. Copy-Number Variations (CNVs) regions in resistance regions of the four biotic stresses analyzed in this study.

List of Supplementary Materials

Supplementary Table 1. Geographic origin and phenotype information of the soybean accessions used in this work.

Supplementary Table 2. Soybean accessions sequencing information

Supplementary Table 3. Variant rate details of the soybeans accessions

Supplementary Table 4. Identified genes with non-synonymous mutations in coding genes of SCN QTLs

Supplementary Table 5. Identified genes with non-synonymous mutations in coding genes of major RKN QTL

Supplementary Figure 1. Summary of the main modification caused by SNPs.

Supplementary Figure 2. QQ-plot for SCN, RKN and SSC analysis.

Supplementary Figure 3. Linkage disequilibrium graphic for SCN race 1 QTL regions. The green lines represent the SNPs related to SCN race 1 resistance.

Supplementary Figure 4. Linkage disequilibrium graphic for *Rhg1* and *Rhg4* regions.

Supplementary Figure 5. Non-synonymous SNPs strongly associated to SCN race 3 resistance on *Rhg1* and *Rhg4* regions of commercial accessions

Supplementary Figure 6. Linkage disequilibrium graphic for *Rdm?* Regions for different intervals

Supplementary Figure 7. Non-synonymous SNPs strongly associated to RKN resistance in major QTL regions of commercial accessions.

Supplementary Figure 8. Copy-Number Variations (CNVs) for commercial cultivar Forrest.

Supplementary Figure 9. Copy-Number Variations (CNVs) regions in SCN QTL regions.

Supplementary Figure 10. Copy-Number Variations (CNVs) for accession HN020.

Supplementary Figure 11. Copy-Number Variations (CNVs) on chromosome 2 for *Rdm1* and *Rdm4* regions.

Table 1. SNPs associated to important disease resistance traits.

Chromosome	Start	End	Trait	Number of SNPs	Type
Chr01	417,636	952,419	SCN Race 1	343	QTL
Chr01	5,579,805	5,579,812	SCN Race 1	2	QTL
Chr07	15,716,384	15,726,458	SCN Race 3	16	QTL
Chr08	7,556,711	7,605,699	SCN Race 1	2	QTL
Chr08	7,918,389	8,368,552	SCN Race 3	502	*Rhg4*
Chr08	8,207,800	8,290,295	SCN Race 1	8	*Rhg4*
Chr08	18,580,944	18,609,706	SCN Race 3	13	QTL
Chr08	18,852,146	18,870,715	SCN Race 1	19	QTL
Chr09	2,078,234	4,506,466	SCN Race 1	32	QTL
Chr09	28,755,071	28,799,989	SCN Race 1	4	QTL
Chr09	32,722,476	32,730,398	SCN Race 1	4	QTL
Chr10	308	1,523,070	RKN	4,461	QTL
Chr10	41,404,597	41,909,795	SCN Race 3	944	QTL
Chr10	43,386,669	43,399,316	SCN Race 3	4	QTL
Chr11	30,651,783	30,660,387	SCN Race 3	19	QTL
Chr11	32,008,820	32,860,835	SCN Race 3	227	QTL
Chr14	1,544,684	2,022,167	SSC	2,014	*Rdm?*
Chr17	339,419	646,272	SCN Race 3	163	QTL
Chr17	13,814,281	23,118,086	SCN Race 1	664	QTL
Chr17	34,753,985	34,901,884	SCN Race 1	5	QTL
Chr18	1,361,675	1,373,360	SCN Race 3	35	QTL
Chr18	1,558,518	1,663,245	SCN Race 3	73	*Rhg1*
Chr18	2,103,323	2,103,361	SCN Race 3	2	QTL
Chr18	3,048,432	3,098,529	SCN Race 3	58	QTL
Chr18	4,358,422	5,136,945	SCN Race 3	24	QTL
Chr18	5,442,907	5,446,608	SCN Race 3	69	QTL
Chr19	40,151,263	40,152,445	SCN Race 1	2	QTL
Chr20	35,025,263	35,395,248	SCN Race 1	378	QTL

Table 2. Summary of regions and type of modifications caused by SNPs for each soybean disease

Type	All	Percentage	SCN Race 1	SCN Race 3	Total	RKN	SSC
Intergenic Region	1,097	10.88%	297	234	531	476	90
3' UTR	311	3.09%	44	85	129	125	57
5' UTR	179	1.78%	24	35	59	83	37
Upstream	5,070	50.30%	631	940	1,571	2,303	1,196
Downstream	1,702	16.89%	233	465	698	755	249
Start codon	0	0.00%	0	0	0	0	0
Start Gained	25	0.25%	3	5	8	11	6
Start Lost	1	0.01%	0	0	0	1	0
Stop Gained	9	0.09%	1	3	4	4	1
Stop Lost	1	0.01%	1	0	1	0	0
Stop Retained	2	0.02%	1	0	1	1	0
Non Synonymous cds	324	3.21%	34	84	118	134	72
Synonymous cds	324	3.21%	46	74	120	135	69
Intron	911	9.04%	133	196	329	383	199
Splice Region	41	0.41%	7	9	16	13	12
Splice Acceptor Region	2	0.02%	0	0	0	1	1
Splice Donor Region	3	0.03%	0	1	1	1	1
none	77	0.76%	7	22	29	24	24
Total	10,079	--	1,462	2,153	3,615	4,450	2,014

Table 3. SNP markers closely associated with the soybean disease resistance traits.

Trait	Chr	Start	End	Number of SNPs	p-value	Minor allele mean	r^2
RKN	10	659,245	707,709	34	1.26E-05	0.4783	0.6385
RKN	10	781,019	952,630	5	1.40E-05	0.2174	0.6318
RKN	10	781,503	811,218	22	3.21E-06	0.2391	0.7298
RKN	10	1,012,047	1,026,489	8	7.78E-06	0.4674	0.6700
RKN	10	1,037,384	1,037,394	2	1.21E-05	0.4891	0.6414
SCN Race 1	1	934,952	952,419	2	1.69E-05	0.3043	0.9529
SCN Race 1	1	937,349	948,132	6	2.05E-05	0.2826	0.9504
SCN Race 1	8	8,290,181	8,290,295	3	5.01E-05	0.3261	0.9394
SCN Race 1	17	13,857,293	17,839,937	141	1.61E-05	0.3043	0.9535
SCN Race 1	20	35,129,880	35,268,944	8	4.37E-05	0.3913	0.9411
SCN Race 3	8	8,209,460	8,368,519	145	5.65E-05	0.3750	0.7011
SCN Race 3	10	41,518,097	41,644,753	4	3.05E-05	0.3646	0.7230
SCN Race 3	18	1,558,518	1,630,473	7	3.44E-05	0.4583	0.7186
SCN Race 3	18	1,621,248	1,631,884	17	3.41E-06	0.4792	0.8062
SCN Race 3	18	1,633,967	1,639,152	8	5.74E-06	0.4896	0.7856
SSC	14	1,700,028	1,713,925	49	1.30E-06	0.4390	0.9050
SSC	14	1,719,777	1,989,255	235	5.48E-07	0.4146	1.0000
SSC	14	1,837,507	1,905,067	2	1.15E-06	0.3780	0.9178
SSC	14	1,896,756	1,932,239	2	6.83E-07	0.4024	0.9754
SSC	14	1,992,573	2,009,713	6	1.28E-06	0.3902	0.9069

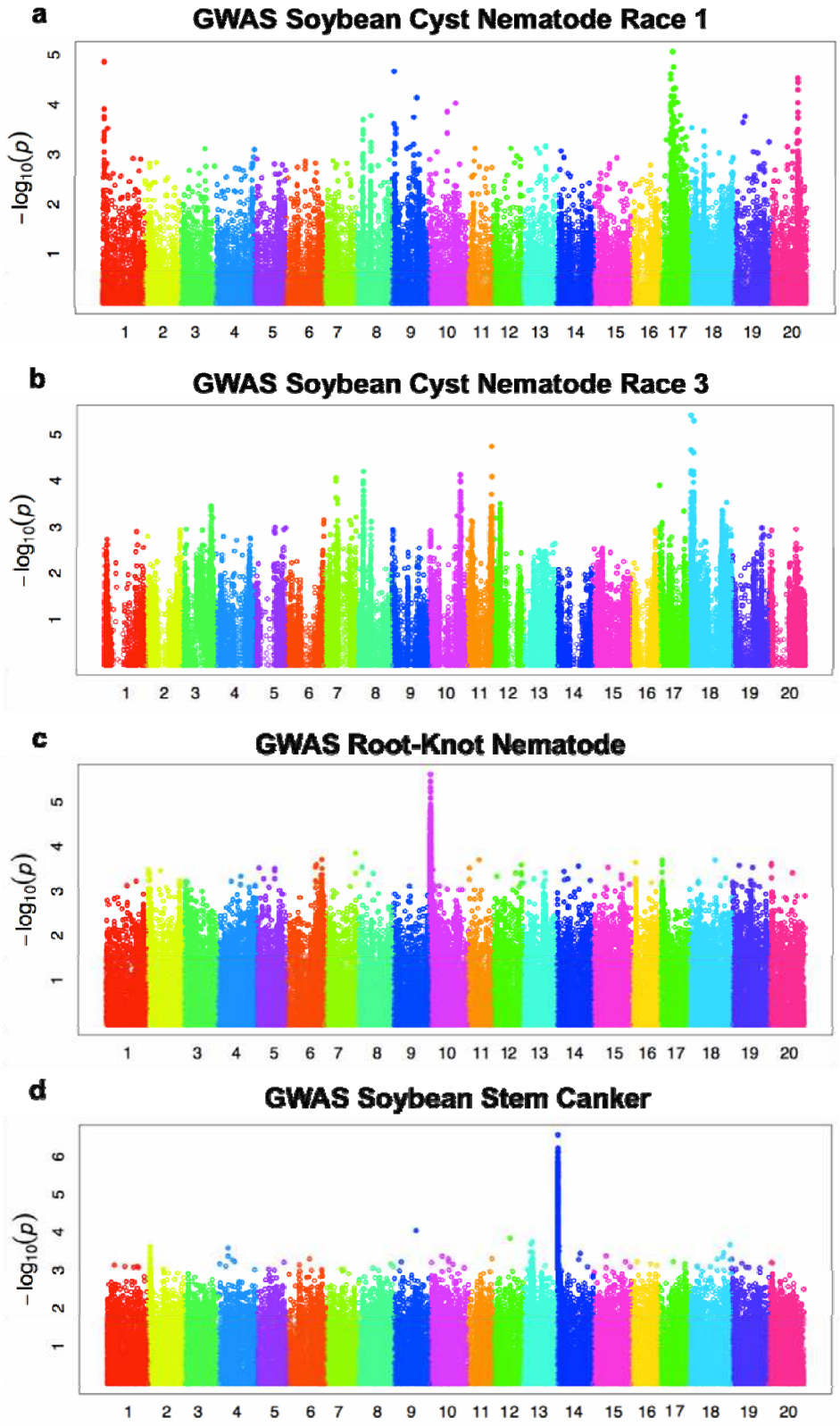


Figure 1. Manhattan plot from the three biotic stresses analyzed in this study. **(a)** SCN race 1 results **(b)** SCN race 3 results, **(c)** RKN results and **(d)** SSC results.

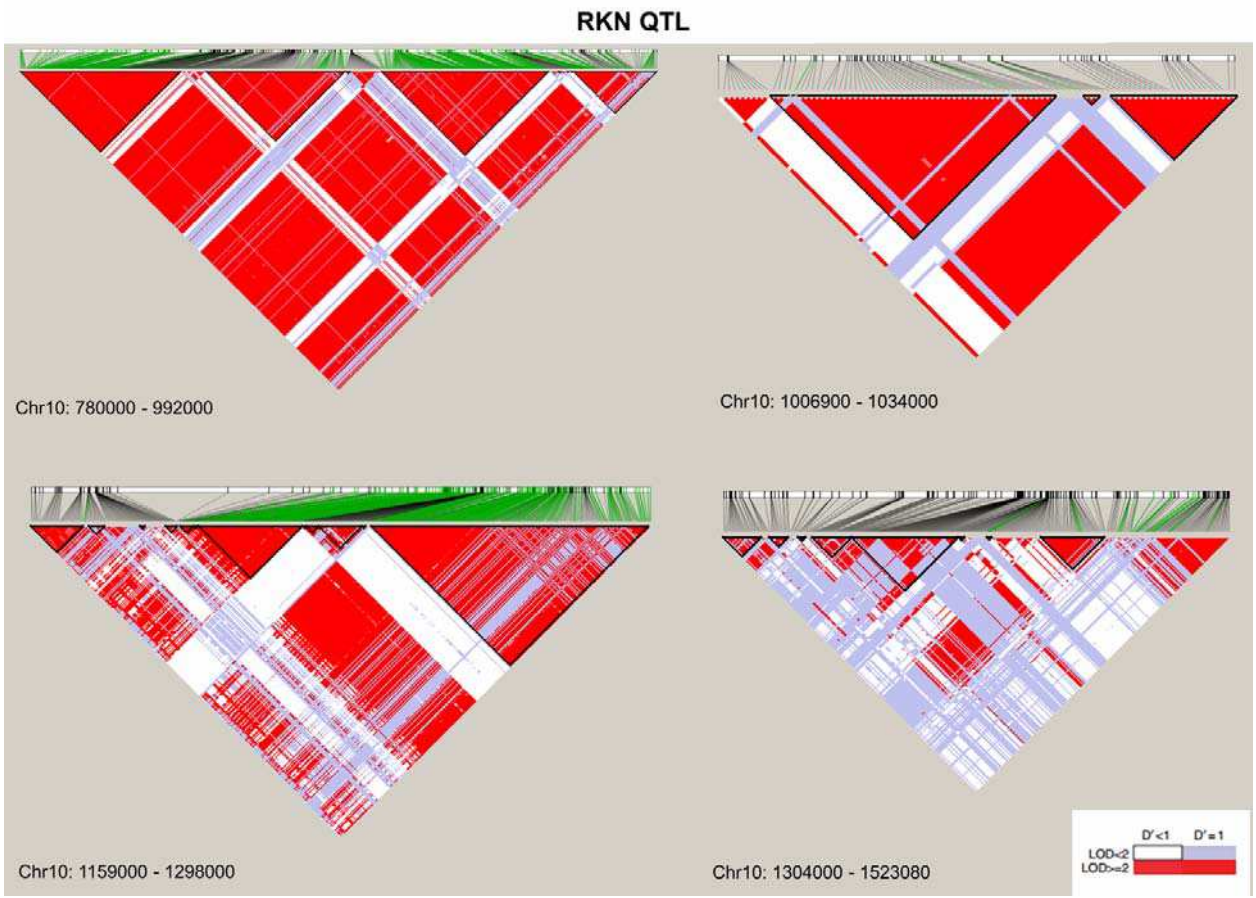


Figure 2. Linkage disequilibrium analysis for RKN QTL regions. The green lines represent the SNPs related to RKN resistance.

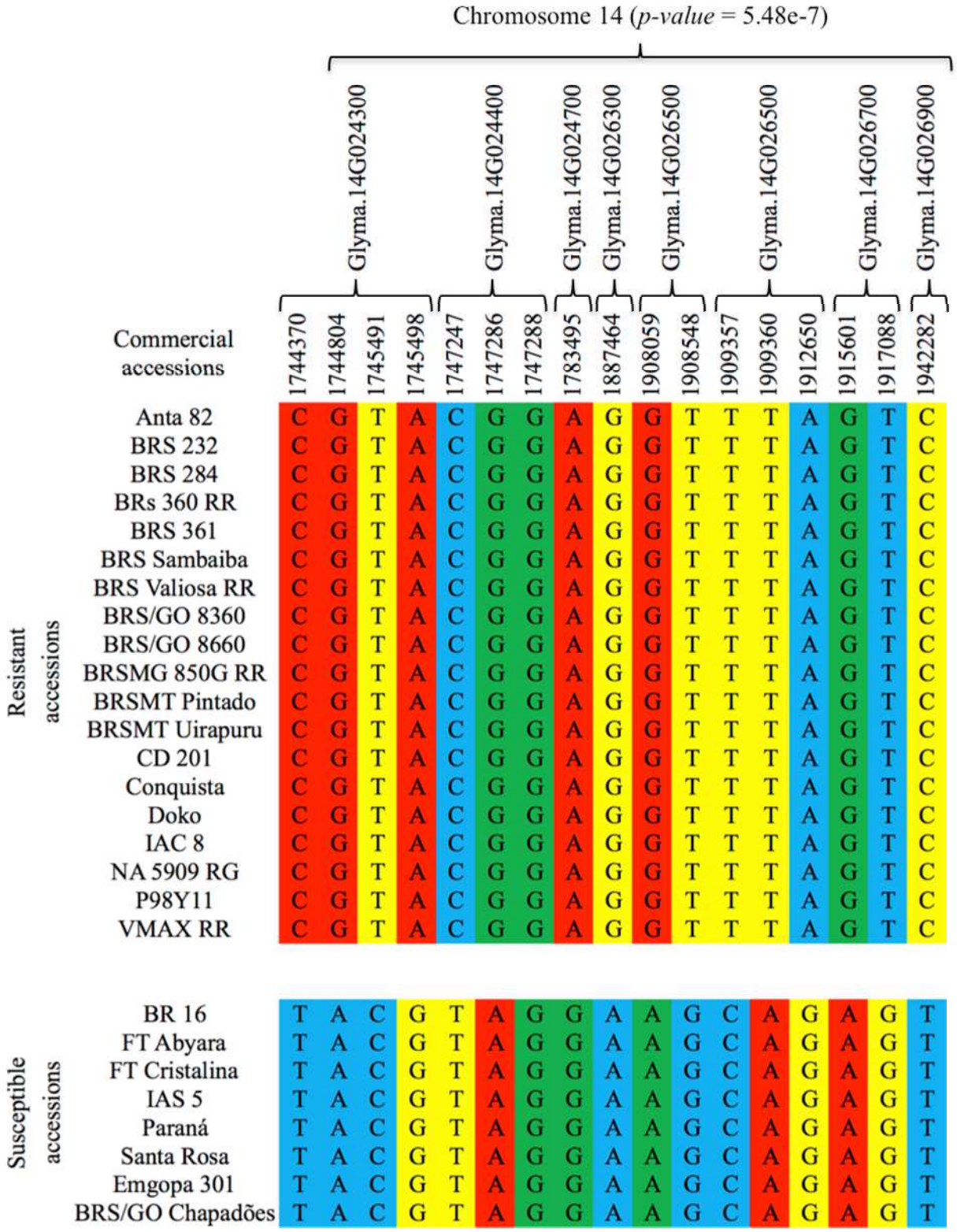


Figure 3. Non-synonymous SNPs haplotypes strongly associated to SSC resistance in *Rdm?* region on commercial accessions.

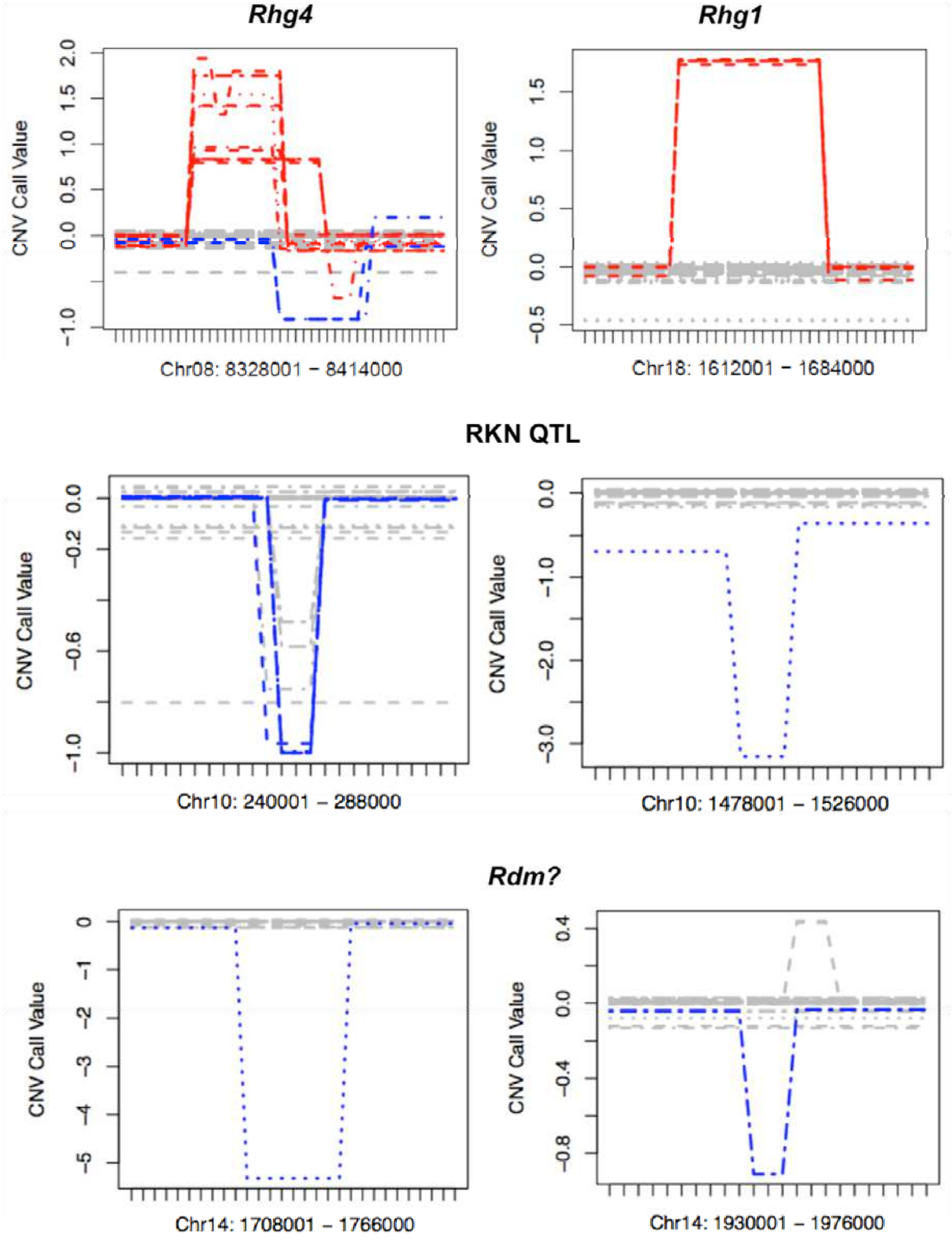


Figure 4. Copy-Number Variations (CNVs) in resistance regions of the four disease resistance traits investigated in this study. The x-axis represents the genomic position and the y-axis the CNV call produced by the segmentation algorithm. The red/blue lines are inserted/deleted fragments detected in these regions.

Supplementary Table 1. Geographic origin and disease phenotype information for the soybean accessions used in this study.

Accession Name	Origin	SCN Race 1	SCN Race 3	RKN	SSC
Anta 82	Brazil	S	R	S	R
BR 16	Brazil	S	S	S	S
BRS 232	Brazil	S	S	R	R
BRS 284	Brazil	S	S	S	R
BRS 360 RR	Brazil	S	S	S	R
BRS 361	Brazil	S	S	S	R
BRS Sambaíba	Brazil	S	S	S	R
BRS Valiosa RR	Brazil	S	S	R	R
BRS/GO 8360	Brazil	S	S	S	R
BRS/GO 8660	Brazil	S	R	R	R
BRS/GO Chapadões	Brazil	R	R	R	S
BRSMG 850G RR	Brazil	S	S	R	R
BRSMT Pintado	Brazil	R	R	S	R
BRSMT Uirapuru	Brazil	S	S	S	R
CD 201	Brazil	S	S	R	R
MG/BR 46 (Conquista)	Brazil	S	S	R	R
Doko	Brazil	S	S	S	R
Emgopa 301	Brazil	S	S	S	MR
FT Abyara	Brazil	S	S	S	S
FT Cristalina	Brazil	S	S	S	S
IAC 8	Brazil	S	S	S	R
IAS 5	Brazil	S	S	S	S
NA 5909 RG	Brazil	S	S	S	R
P98Y11	Brazil	R	R	S	R
Paraná	Brazil	S	S	S	S
Santa Rosa	Brazil	S	S	S	S
VMAX RR	Brazil	S	R	S	R
Forrest	United States	R	R	R	S
G93-9223	United States	NA	R	R	MR
Magellan	United States	S	S	S	NA
Williams 82	United States	S	S	S	NA
HN001	United States	S	S	NA	NA
HN002	China	R	R	NA	S
HN003	China	R	R	S	R
HN004	China	R	R	S	S
HN005	China	R	R	S	NA
HN006	Korea	NA	NA	S	NA
HN007	Korea	NA	NA	S	NA
HN008	China	R	R	NA	NA
HN009	China	NA	NA	S	NA
HN010	China	R	R	NA	S
HN011	Japan	R	R	NA	NA
HN012	China	NA	NA	R	S
HN014	United States	S	S	NA	S
HN015	China	R	R	S	NA
HN016	China	S	S	R	R
HN018	China	R	R	S	NA
HN020	China	S	R	S	R
HN021	Japan	MS	R	R	R
HN022	China	R	R	R	S
HN023	Korea	NA	S	NA	S
HN025	China	R	MR	R	R
HN026	Korea	NA	NA	S	NA

R, Resistant; **MR**, Moderately Resistant; **MS**, Moderately Susceptible**; S**, Susceptible; **NA**, Not Available

Supplementary Table 2. Soybean accessions sequencing information

Name	Number of reads	Number of mapped reads	Genome coverage	Mean depth
Anta 82	195,886,578	180,626,491	0.9221	16.7789
BR 16	196,397,054	186,322,940	0.9487	15.2675
BRS 232	122,368,531	116,984,169	0.9560	9.5324
BRS 284	171,461,978	163,468,763	0.9534	13.0724
BRS 360 RR	201,402,002	192,682,206	0.9567	15.8585
BRS 361	191,425,927	183,136,485	0.9567	15.0523
BRS Sambaíba	192,188,718	181,424,387	0.9440	15.2387
BRS Valiosa RR	156,259,990	148,104,258	0.9478	11.8584
BRS/GO 8360	158,327,158	148,939,207	0.9407	13.6017
BRS/GO 8660	142,045,598	133,524,801	0.9400	10.8659
BRS/GO Chapadões	245,514,622	222,293,654	0.9054	12.5297
BRSMG 850G RR	217,891,043	206,848,122	0.9493	16.9610
BRSMT Pintado	191,564,775	181,728,921	0.9487	14.9496
BRSMT Uirapuru	198,605,107	186,389,164	0.9385	11.4300
CD 201	221,562,422	210,957,959	0.9521	17.6887
Conquista	272,124,634	258,656,049	0.9505	11.1745
Doko	247,331,921	234,203,074	0.9469	15.6906
Emgopa 301	159,411,436	150,538,085	0.9443	19.7702
FT Abyara	185,414,295	175,453,981	0.9463	16.4883
FT Cristalina	223,191,645	213,324,477	0.9558	14.3955
IAC 8	204,515,979	194,690,635	0.9520	18.1685
IAS 5	150,707,920	133,762,039	0.8876	16.2044
NA 5909 RG	189,097,655	179,350,133	0.9485	22.0393
P98Y11	149,136,842	141,849,489	0.9511	14.2505
Paraná	317,943,923	301,616,213	0.9486	11.5994
Santa Rosa	166,171,636	157,804,753	0.9496	24.6117
VMAX RR	221,332,490	207,968,442	0.9396	13.0282
HN025	162,542,041	147,802,280	0.9093	10.2156
HN009	188,809,910	166,874,460	0.8838	11.3138
HN015	195,357,968	171,712,261	0.8790	11.1497
HN008	165,486,559	144,376,365	0.8724	13.3146
HN012	161,132,105	143,472,128	0.8904	10.6777
HN016	152,787,788	137,652,993	0.9009	13.4266
HN005	200,310,937	180,267,479	0.8999	11.8758
HN002	189,724,622	169,489,761	0.8933	11.9288
HN020	132,029,140	116,428,620	0.8818	12.1235
HN003	197,816,271	177,036,382	0.8950	13.4241
HN004	175,091,860	155,245,356	0.8867	13.9184
HN010	149,642,173	134,821,340	0.9010	12.9727
HN018	174,487,929	155,331,036	0.8902	10.6026
HN022	179,241,390	162,441,938	0.9063	13.1667
HN021	189,123,705	171,425,013	0.9064	8.9867
HN011	161,362,060	130,592,691	0.8093	10.2415
HN026	182,522,344	163,101,808	0.8936	11.1124
HN007	195,461,329	175,512,165	0.8979	13.0076
HN023	151,462,168	137,522,528	0.9080	12.4005
HN006	185,556,824	168,649,278	0.9089	14.1835
Forrest	173,011,062	138,493,887	0.8005	12.3882
HN014	167,020,416	148,757,679	0.8907	11.0623
G93-9223	258,642,533	245,092,630	0.9476	10.2718
HN001	178,011,768	158,068,790	0.8880	18.3659
Magellan	177,903,577	158,768,370	0.8924	20.5986

Supplementary Table 3. Variant rate details of the soybeans accessions

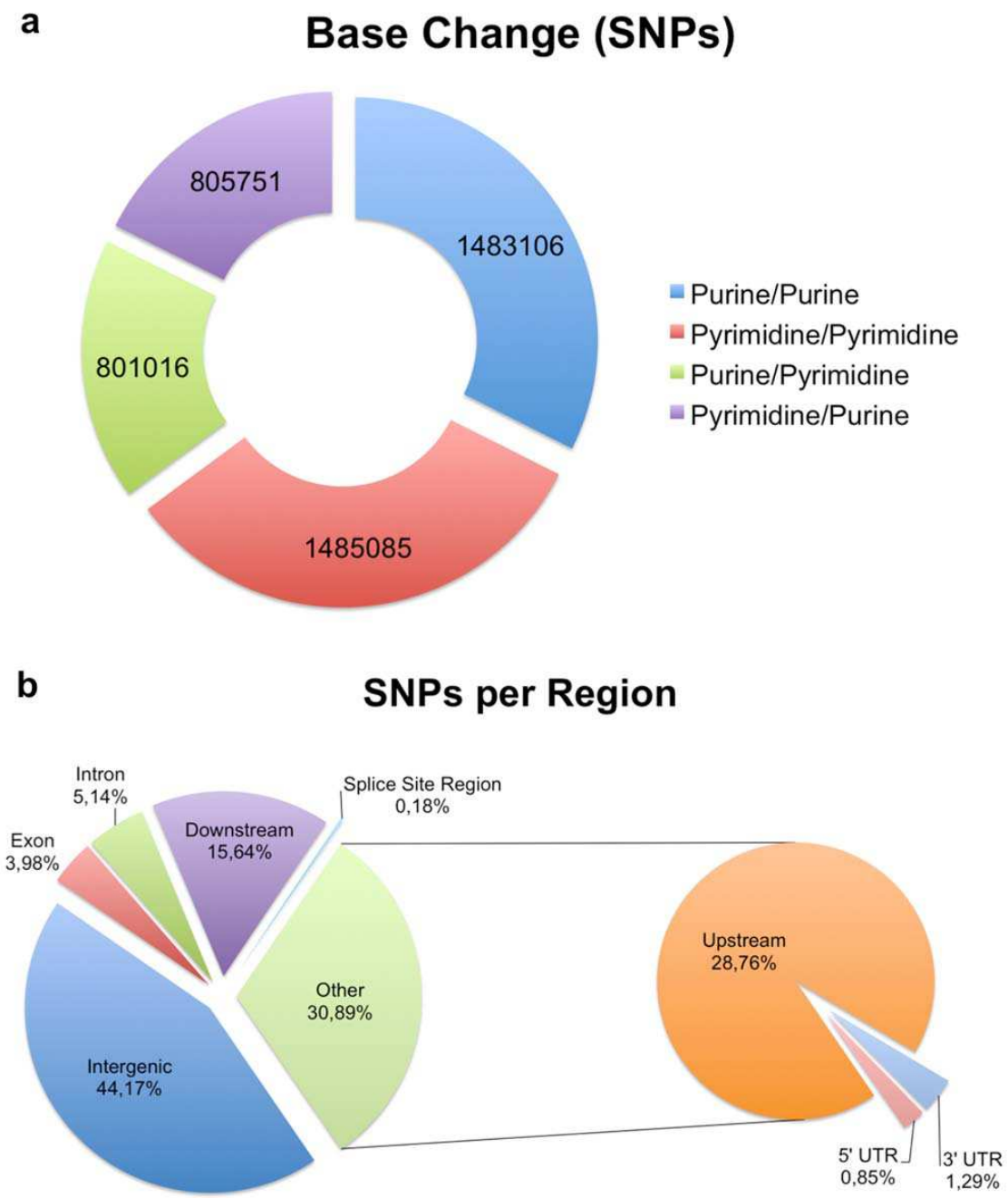
Chromosome	Length	Variants	Variants rate (SNP per bp)
Chr01	56,831,624	202,160	281
Chr02	48,577,505	223,611	217
Chr03	45,779,781	232,228	197
Chr04	52,389,146	248,168	211
Chr05	42,234,498	137,400	307
Chr06	51,416,486	252,399	204
Chr07	44,630,646	184,814	241
Chr08	47,837,940	181,497	264
Chr09	50,189,764	249,431	201
Chr10	51,566,898	171,412	301
Chr11	34,766,867	124,008	280
Chr12	40,091,314	145,714	275
Chr13	45,874,162	242,791	189
Chr14	49,042,192	234,327	209
Chr15	51,756,343	356,197	145
Chr16	37,887,014	233,871	162
Chr17	41,641,366	209,361	199
Chr18	58,018,742	424,845	137
Chr19	50,746,916	247,917	205
Chr20	47,904,181	209,599	229

Supplementary Table 4. Genes with non-synonymous mutations in coding sequences on SCN QTLs regions.

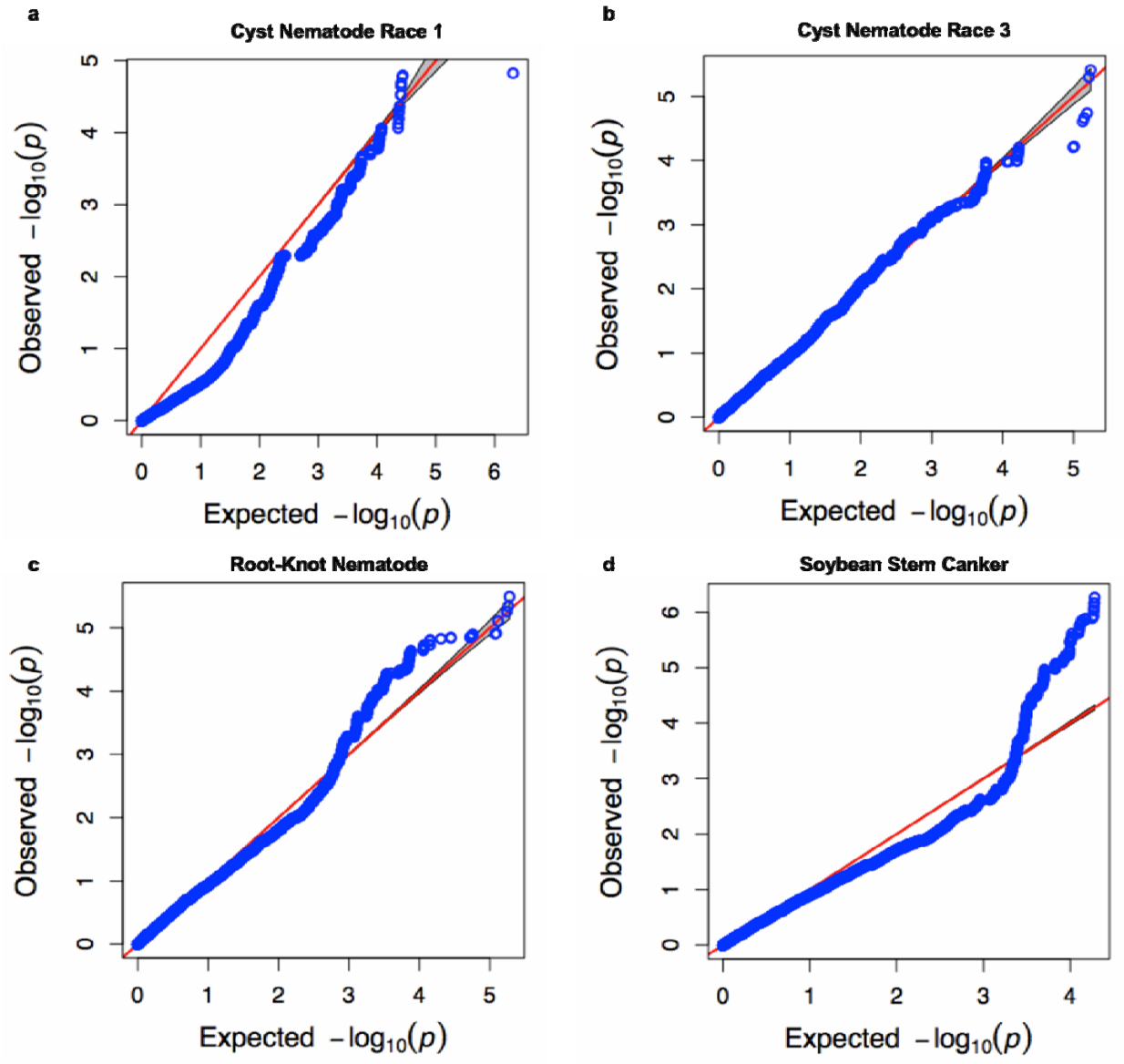
Description	Gene Name
5'-AMP-activated protein kinase	*Glyma.17G006000*
Acetylglutamate kinase/synthase)	*Glyma.17G160300*
Calcium transporting ATPase	*Glyma.17G161300*
Casein kinase	*Glyma.10G181700*
Chaperone-activity of BC1 complex [CABC1]-related	*Glyma.18G028000*
Cyclin	*Glyma.17G160000, Glyma.17G175700 and Glyma.20G107700*
Cytochrome P450	*Glyma.10G202400, and Glyma.17G007200*
Myb-like DNA-binding domain	*Glyma.01G004900, Glyma.01G009600, Glyma.10G180800 and Glyma.20G108600*
NADH-ubiquinone oxidoreductases	*Glyma.08G228300, Glyma.17G160400 and Glyma.20G108500*
Nipped-b-like protein [delangin] SCC2-related)	*Glyma.20G107600*
O-methyltransferase	*Glyma.01G004200*
Pathogenesis-related protein Bet v I family related to response to biotic stimulus	*Glyma.08G230400 and Glyma.08G230500*
Poly[A]-specific exoribonuclease PARN related to RNA binding	*Glyma.10G185400*
Polyketide cyclases related to response to biotic stimulus	*Glyma.01G009200 and Glyma.11G233300.*
PPR repeat domain	*Glyma.01G004400, Glyma.17G160600 and Glyma.20G109800*
pre-mRNA-splicing factor SYF2	*Glyma.20G107700*
Prenylated rab acceptor 1	*Glyma.10G180400*
Protein phosphatase 2 regulatory subunit	*Glyma.10G181800*
RNA-directed RNA polymerase QDE-1 required for post-transcriptional gene silencing and RNA interference	*Glyma.01G008700*
Serine-threonine kinase	*Glyma.08G099400, Glyma.10G181500, Glyma.10G181700, Glyma.10G186000, Glyma.11G230300, Glyma.17G173000 and Glyma.20G109400*
THUMP domain-containing proteins related to RNA binding	*Glyma.20G108700*
Transcription factors	*Glyma.11G227800, Glyma.11G231300 and Glyma.17G005600*
Zinc finger, C3HC4 type domain	*Glyma.01G009100, Glyma.11G231400, Glyma.17G160500 and Glyma.17G172200*

Supplementary Table 5. Genes with non-synonymous mutations in coding sequences on major RKN QTL regions.

Description	Gene
AMP-binding enzyme	Glyma.10G0
AP2 domain related to regulation of transcription	Glyma.10G007100 and G
BRG-1 associated factor 250 [BAF250]	Glyma.10G0
Chlorophyllase	Glyma.10G0
CIRCADIAN PROTEIN CLOCK/ARNT/BMAL/PAS	Glyma.10G0
Cytochrome b5-like Heme/Steroid binding domain	Glyma.10G0
DNA helicases	Glyma.10G000200 and G
DNA REPLICATION REGULATOR DPB11-RELATED	Glyma.10G0
Enoyl-CoA hydratase/isomerase	Glyma.10G0
Glycosyl hydrolase	Glyma.10G013900, Glyma.10G017
Heat shock transcription factor	Glyma.10G0
Iron/ascorbate family oxidoreductases	Glyma.10G0
LATE EMBRYOGENESIS ABUNDANT [PLANTS] LEA-RELATED	Glyma.10G0
Lipase [class 3]	Glyma.10G0
Molecular chaperone	Glyma.10G0
Myb-like DNA-binding domain	Glyma.10G0
MYOSIN	Glyma.10G0
NADH-ubiquinone oxidoreductase flavoprotein 1	Glyma.10G0
Nucleolar protein involved in 40S ribosome biogenesis	Glyma.10G0
PPR repeat domain	Glyma.10G001500, Glyma.10G011
Predicted small molecule transporter	Glyma.10G0
REGULATOR OF CHROMOSOME CONDENSATION with FYVE zinc finger	Glyma.10G0
serine-type peptidase activity	Glyma.10G0
Serine/threonine protein kinase	Glyma.10G0
Splicing factor U2AF, large subunit [RRM superfamily]	Glyma.10G0
Transcription factor	Glyma.10G0
Transferase family	Glyma.10G0
Translation initiation factor	Glyma.10G001300 and G
tRNAHis guanylyltransferase related to tRNA modification	Glyma.10G0
VACUOLAR PROTEIN SORTING-ASSOCIATED PROTEIN 51 HOMOLOG	Glyma.10G0
WD domain, G-beta repeat	Glyma.10G0
Zinc finger domain	Glyma.10G007200, Glyma.10G00

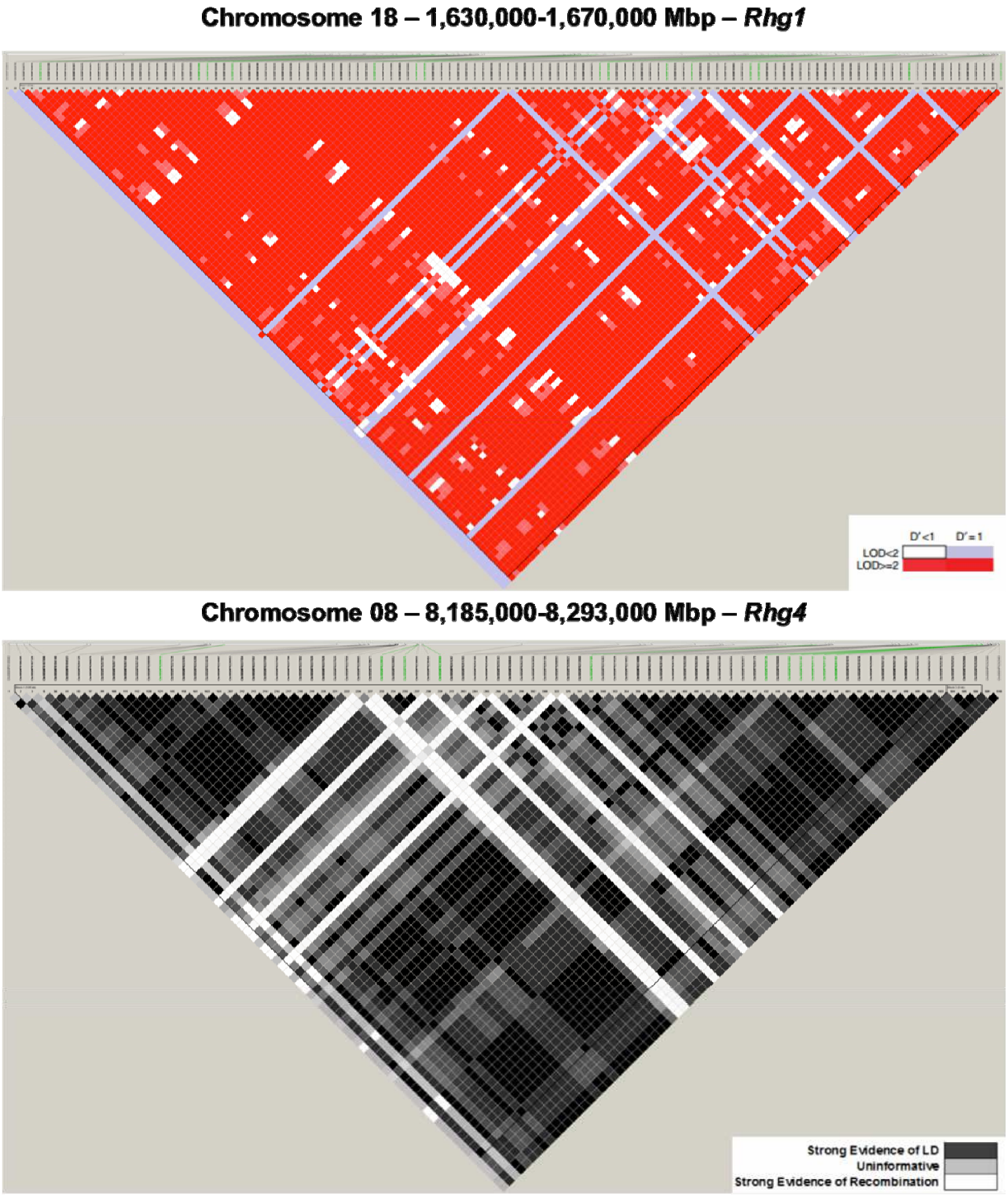


Supplementary Figure 1. Summary of the main modification caused by SNPs. (**a**) Number of transition/transversion mutations (**b**) Percentage of SNPs per region of the soybean genome

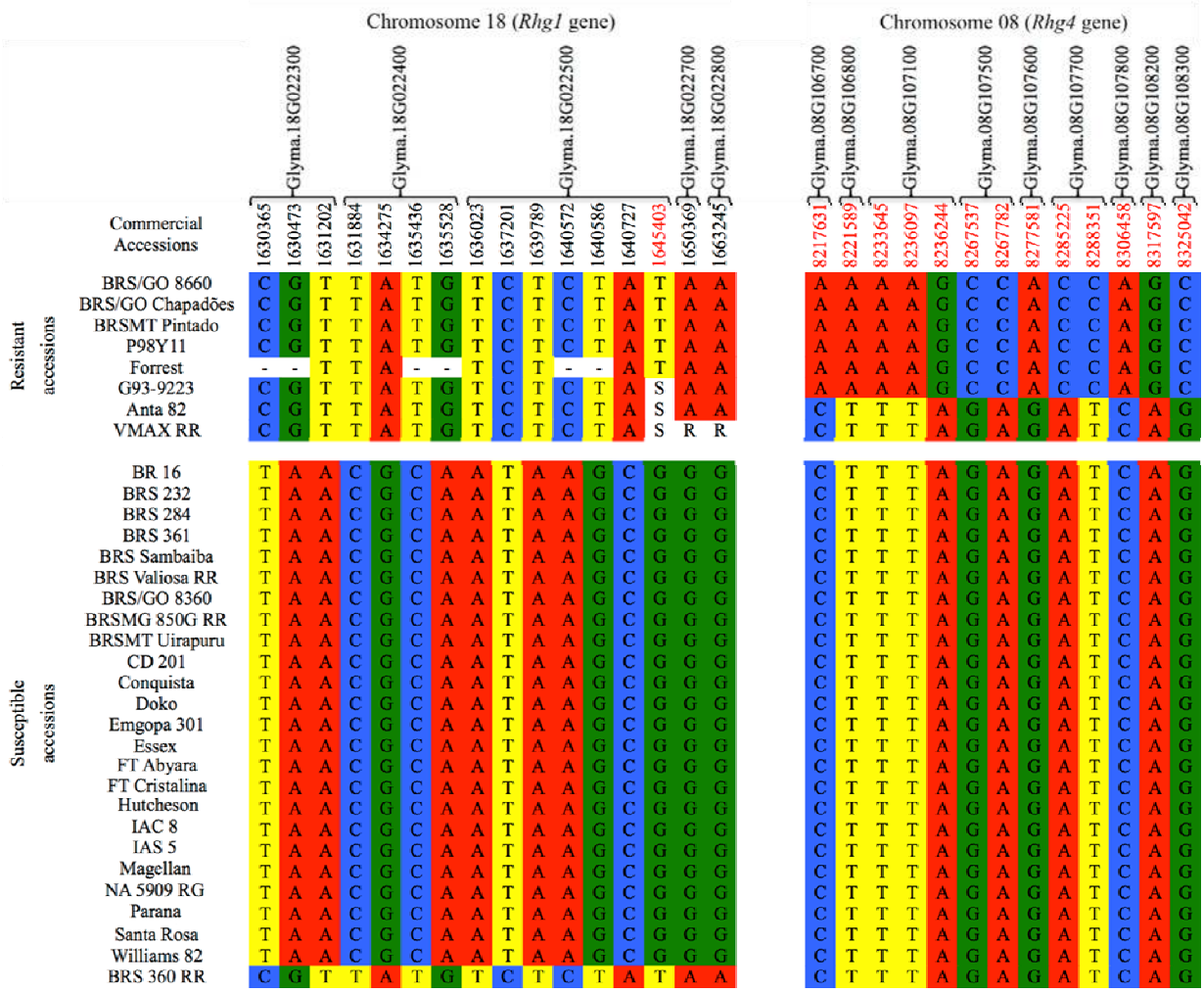


Supplementary Figure 2. QQ-plot analysis for SCN, RKN and SSC resistance traits.

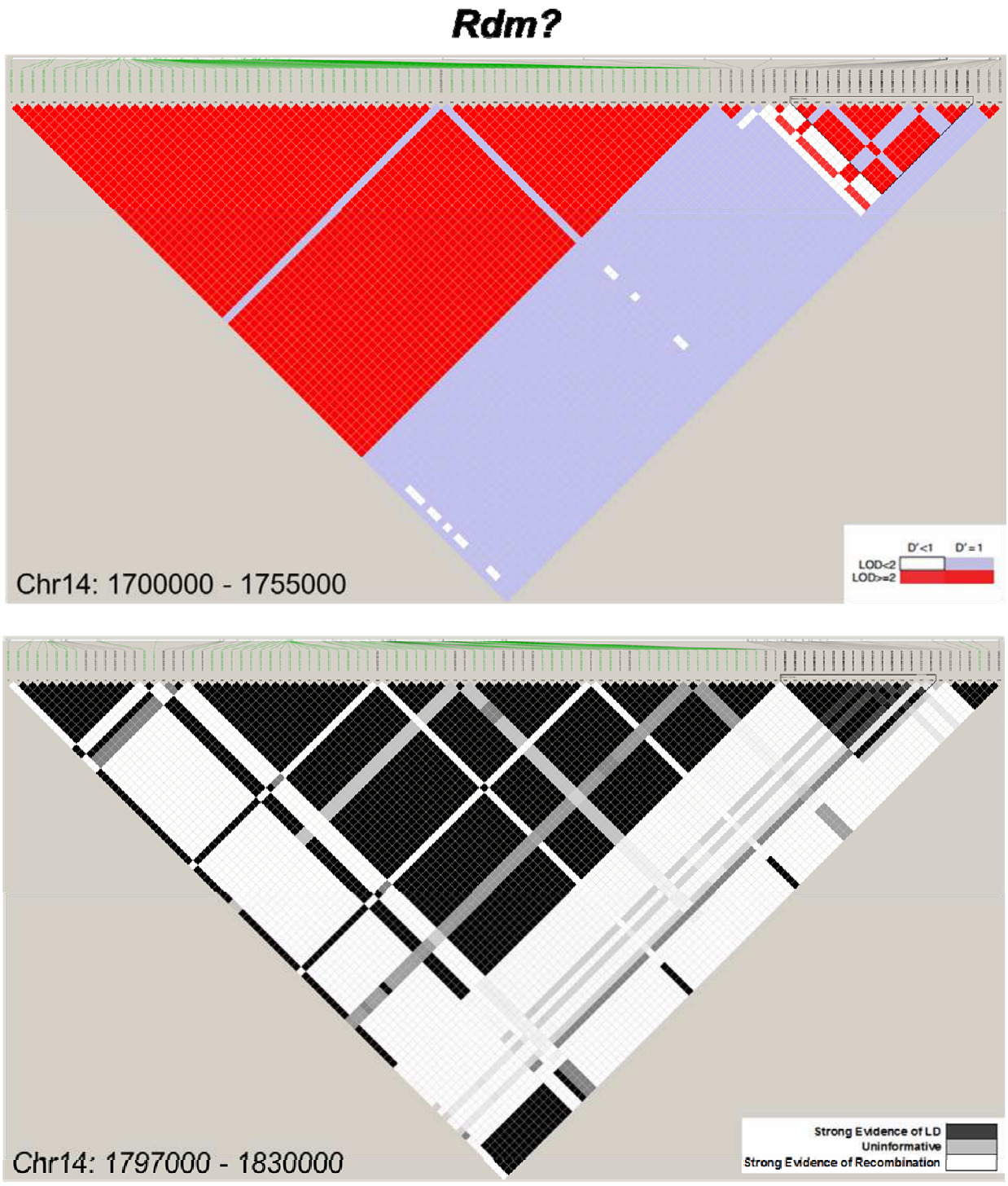
Supplementary Figure 3. Linkage disequilibrium graphic for SCN race 1 QTL regions. The green lines represent the SNPs related to SCN race 1 resistance.



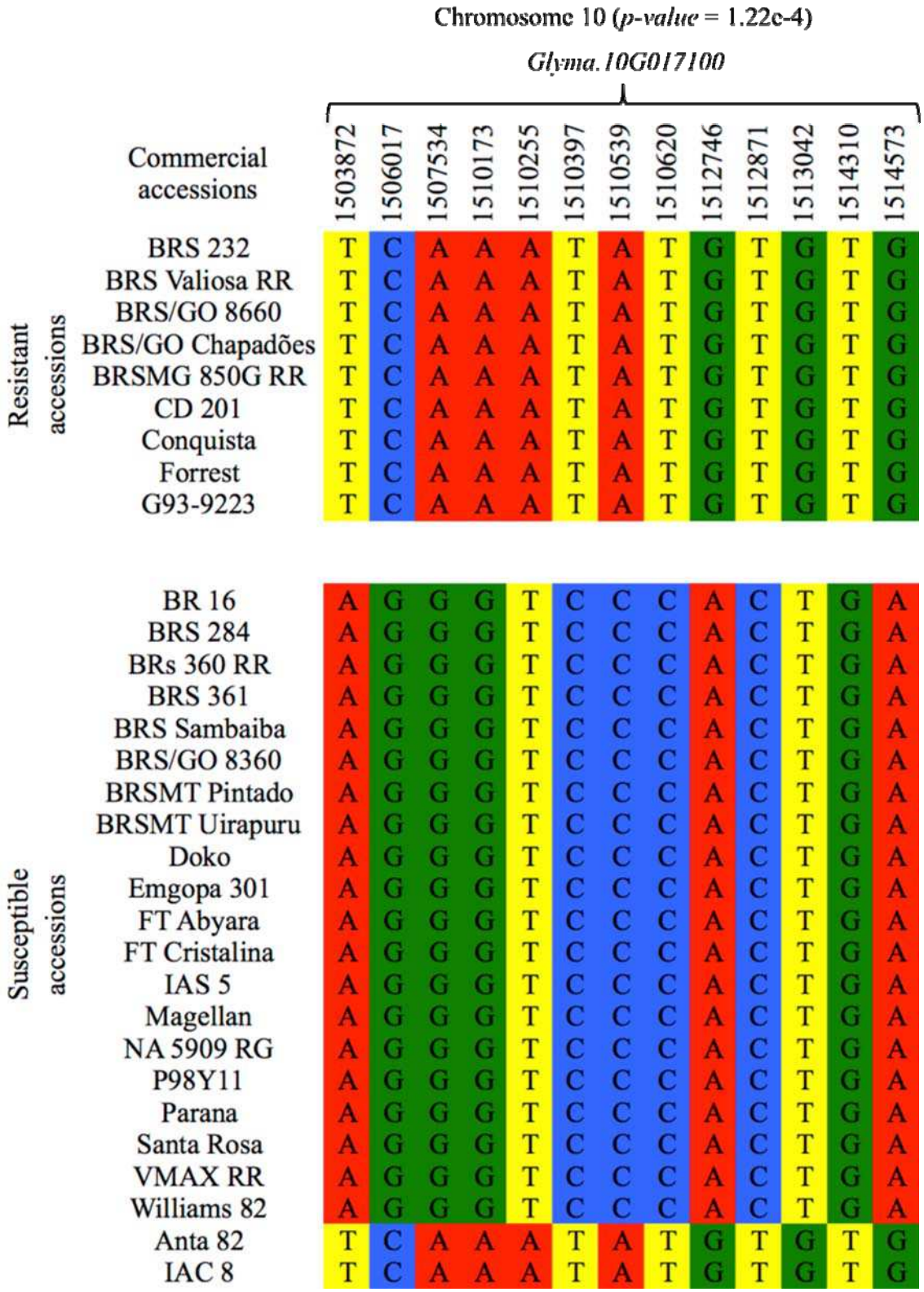
Supplementary Figure 4. Linkage disequilibrium graphic for *Rhg1* and *Rhg4* regions. The first graphic is *Rhg1* region according to D' information data. The second figure is *Rhg4* region according to r^2 information data. The green lines represent the SNPs related to SCN resistance.



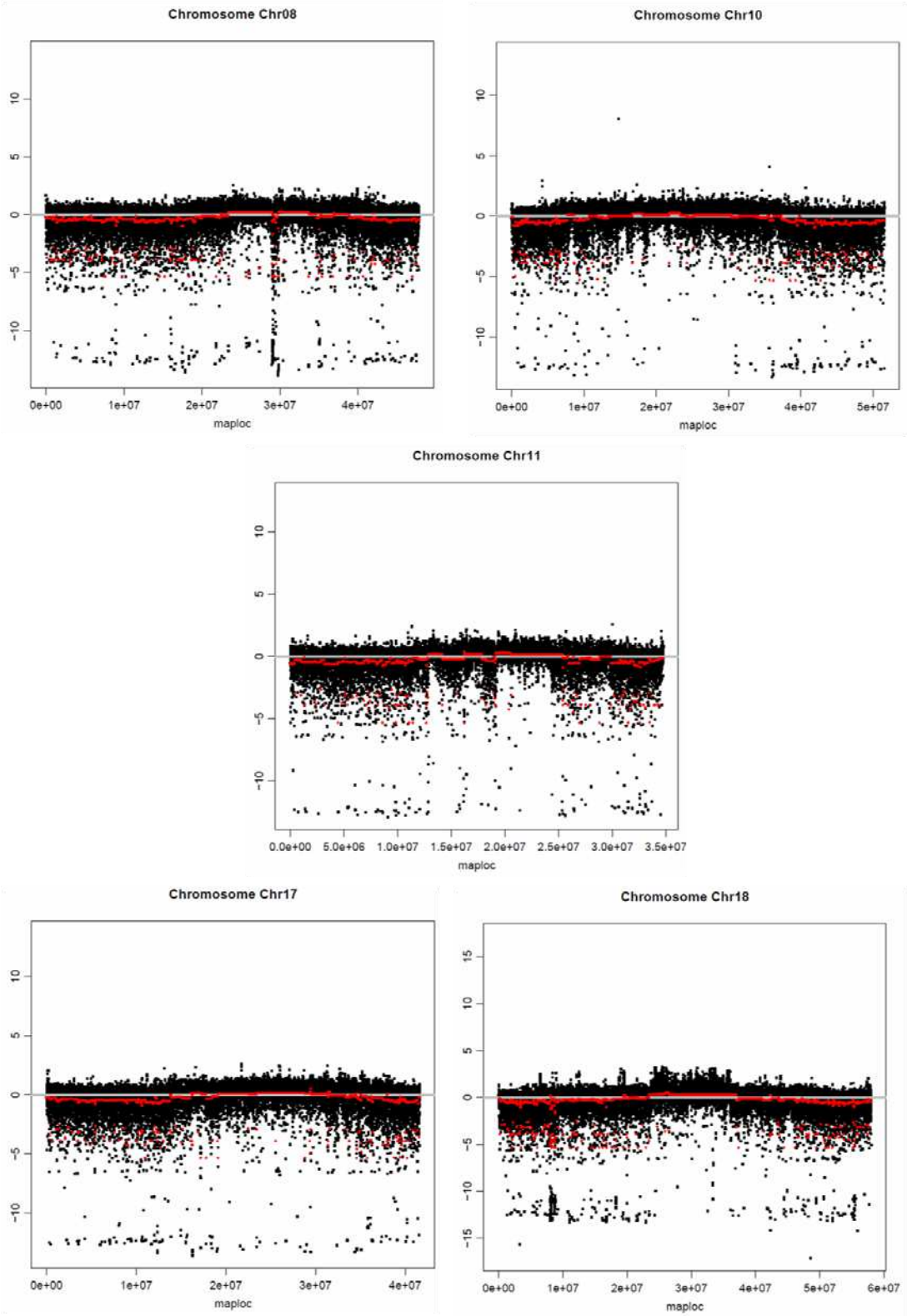
Supplementary Figure 5. Non-synonymous SNPs strongly associated to SCN race 3 resistance in *Rhg1* and *Rhg4* regions of commercial accessions.



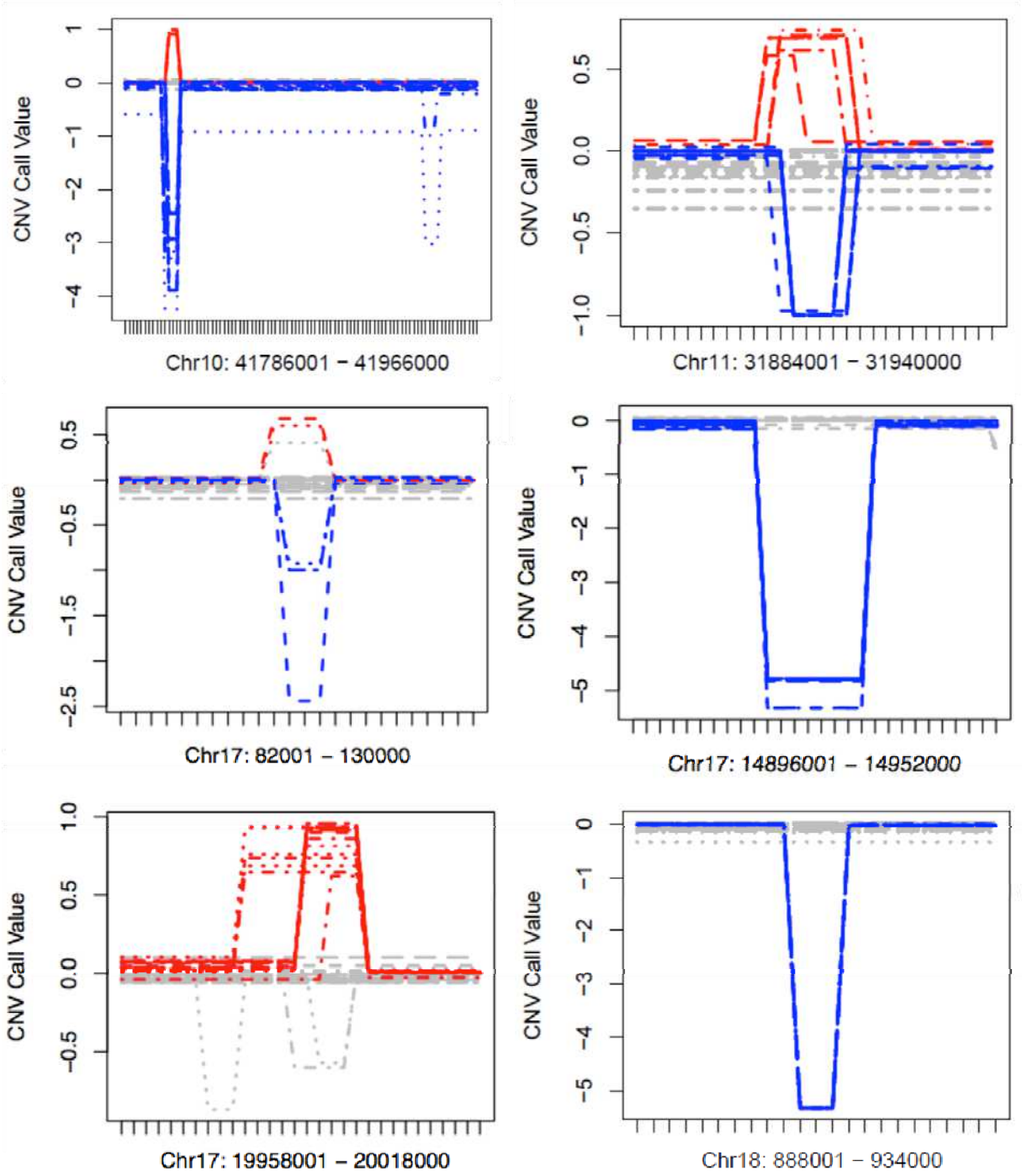
Supplementary Figure 6. Linkage disequilibrium graphic for *Rdm?* region for different intervals. The first graphic is one interval of *Rdm?* gene according to D' information data. The second figure is another interval of *Rdm?* gene according to r^2 information data. The green lines represent the SNPs related to SSC resistance.



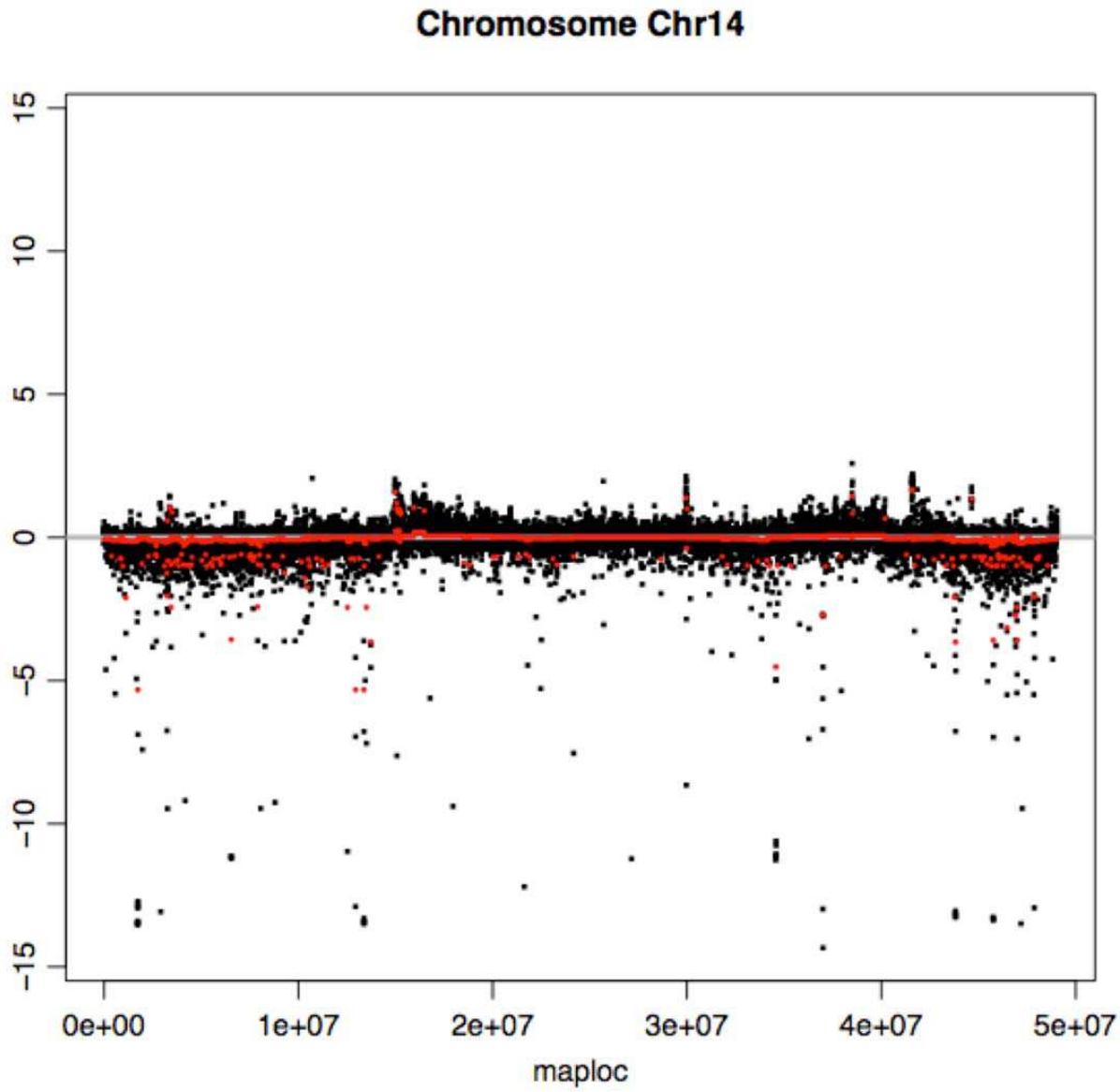
Supplementary Figure 7. Non-synonymous SNPs haplotypes closely associated to RKN resistance in major QTL regions of commercial soybean accessions.



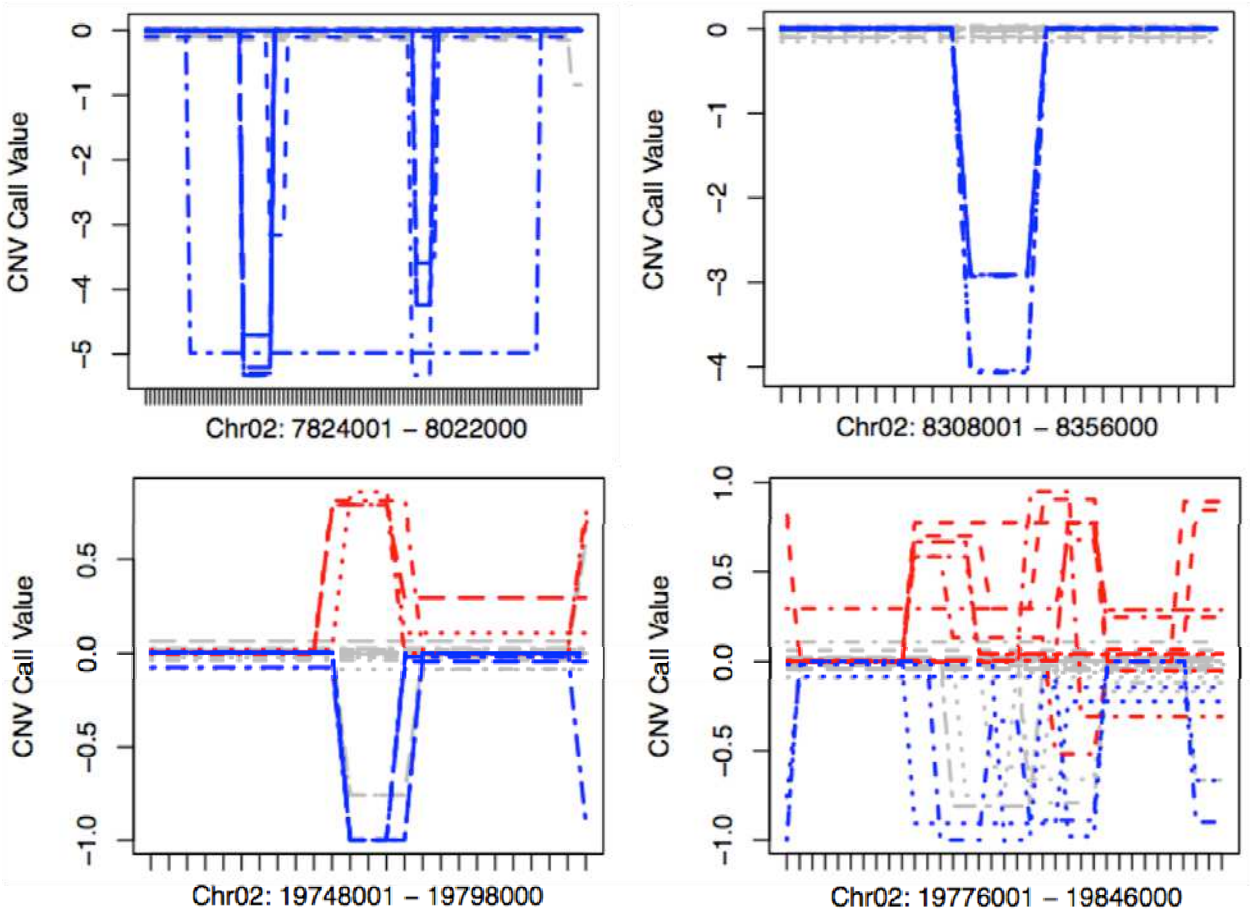
Supplementary Figure 8. Copy-Number Variations (CNVs) for cultivar Forrest. The x-axis represents the genomic position and y-axis the log-ratio of the read counts. The red dots are the copy number call of each segment.



Supplementary Figure 9. Copy-Number Variations (CNVs) in SCN QTL regions. The x-axis represents the genomic position and the y-axis the CNV call produced by the segmentation algorithm. The red/blue lines are inserted/deleted fragments detected in these regions.



Supplementary Figure 10. Copy-Number Variations (CNVs) for accession HN020. The x-axis represents the genomic position and y-axis the log-ratio of the read counts. The red dots are the copy number call of each segment.



Supplementary Figure 11. Copy-Number Variations (CNVs) on chromosome 2 for *Rdm1* and *Rdm4* regions. The x-axis represents the genomic position and the y-axis the CNV call produced by the segmentation algorithm. The red/blue lines are inserted/deleted fragments detected in these regions.

8. CONSIDERAÇÕES FINAIS

Os resultados obtidos nesse estudo podem apresentar grande impacto em programas de melhoramento genético da soja brasileira. Conclui-se que a maior parte das modificações alélicas observadas ao longo dos 50 anos de história da cultura no país encontra-se em regiões não codantes, mais especificamente em regiões intergênicas. Em regiões codantes, existe uma grande quantidade de mutações não sinônimas compartilhadas em todas as cultivares nacionais, ligadas a genes envolvidos com processos de morte celular, fotossíntese, geração de precursores de metabólitos e energia. Dentre os cultivares com a maior presença de SNPs diferentes do genoma referência, encontram-se Doko e Santa Rosa. Além disto, cultivares mais recentes como Anta 82 e VMAX RR apresentam a menor quantidade de variações alélicas em comparação ao genoma referência da soja.

Os dados gerados neste trabalho ainda sugerem que a base genética brasileira da soja ainda continua homogênea, estreita e muito próxima à base genética americana. Tal constatação aumenta a necessidade de introdução de germoplasma exóticos para cruzamento com cultivares nacionais com objetivo de aumentar a diversidade da base genética brasileira. Contudo, existem indícios de um processo de diversificação nos acessos modernos, além da presença de regiões sob processo de seleção positiva, principalmente no cromossomo 17.

Além disto, ressalta-se que foram identificadas 10.079 variações alélicas relacionadas a mecanismos de defesa contra SCN, RKN e SSC. Para resistência contra SCN raça 1, foram observados SNPs inseridos dentro do gene *Rhg4* e em QTLs do cromossomo 1, 8, 9, 17, 19 e 20. As marcas mais significativas encontram-se no cromossomo 17. Já para SCN da raça 3, foram encontrados SNPs inseridos dentro dos genes *Rhg1*, *Rhg4* e em QTLs dos cromossomos 7, 8, 10, 11, 17 e 18. O gene *Rhg1* apresentou os SNPs mais significativos em relação à SCN raça 3. Além disto, o cromossomo 11 apresenta um QTL com muita influência sobre os acessos comerciais brasileiras, podendo ser uma importante fonte de resistência à doença. Conclui-se por estas análises que o QTL localizado no cromossomo 10 é a principal fonte de resistência dos acessos estudados contra RKN. Ainda, através do trabalho foi possível concluir que o gene *Rdm?* representa a principal fonte de resistência contra SSC causado por *Diaporthe phaseolorum f. sp. meridionalis.* Uma grande

quantidade de regiões em alto desequilíbrio de ligação foi observada entre os acessos resistentes às três doenças, sendo a região correspondente ao QTL de resistência a RKN com a maior quantidade delas, com blocos de tamanho variado. Ainda, observou-se a presença de inúmeras mutações não sinônimas em regiões importantes de genes de defesa da planta, como proteínas com domínios MYB, LRR e PPR, fatores de transcrição e proteínas de choque térmico, serinas-treoninas quinases e proteínas relacionadas à patogênese. Tais variações alélicas observadas nestes genes podem desempenhar um papel fundamental nos mecanismos de defesa da planta contra ação dos patógenos. Os resultados observados para esta análise possuem impacto importante em programas de seleção assistida por marcadores para genotipagem e escolha de acessos com resistência contra os três fatores bióticos. Contudo, devido ao tamanho da amostragem utilizada neste trabalho, um processo de validação será necessário para confirmação dos resultados obtidos via GWAS.

Por fim, foi encontrada uma grande quantidade de grandes inserções (CNVs) nas cultivares modernas no cromossomo 16. Outros cromossomos também apresentam CNVs que distinguem cultivares antigas das modernas, com a presença de introgressões e deleções por todo genoma da soja. Além disto, foi possível observar, nos cromossomos 3, 4, 5, 6, 9 e 14, a existência de oito sub-regiões contendo CNVs capazes de distinguir o germoplasma brasileiro do americano. Tais modificações estruturais podem ser peças-chave para explicar as modificações significativas observadas na história de produção e adaptabilidade do germoplasma brasileiro, bem como as diferenças existentes entre os germoplasma brasileiros e americanos. Além disto, os CNVs diferenciadores entre acessos resistentes e suscetíveis detectados em regiões que possuem mecanismos de resistência genética contra os três fatores bióticos podem possuir importante papel na planta. Assim, estudos mais aprofundados sobre estas regiões de CNVs serão fundamentais futuramente para identificação da importância biológica de tais modificações estruturais.