IDENTIFICAÇÃO E PRIORIZAÇÃO DE GENES DE RESISTÊNCIA A ESTRESSES BIÓTICOS EM SOJA (*Glycine max* L. Merr.) A PARTIR DA INTEGRAÇÃO DE ASSOCIAÇÃO GENÔMICA AMPLA E REDES DE COEXPRESSÃO GÊNICA

FABRÍCIO DE ALMEIDA SILVA

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE
DARCY RIBEIRO

CAMPOS DOS GOYTACAZES – RJ
JANEIRO - 2022

IDENTIFICAÇÃO E PRIORIZAÇÃO DE GENES DE RESISTÊNCIA A ESTRESSES BIÓTICOS EM SOJA (*Glycine max* L. Merr.) A PARTIR DA INTEGRAÇÃO DE ASSOCIAÇÃO GENÔMICA AMPLA E REDES DE COEXPRESSÃO GÊNICA

FABRÍCIO DE ALMEIDA SILVA

Dissertação apresentada ao Centro de Biociências e Biotecnologia da Universidade Estadual do Norte Fluminense Darcy Ribeiro, como parte das exigências para obtenção do título de Mestre em Biotecnologia Vegetal.

Orientador: Dr. Thiago Motta Venancio

CAMPOS DOS GOYTACAZES – RJ
JANEIRO – 2022

FABRICIO DE ALMEIDA SILVA

IDENTIFICAÇÃO E PRIORIZAÇÃO DE GENES DE RESISTÊNCIA A ESTRESSES BIÓTICOS EM SOJA (*Glycine max* L. Merr.) A PARTIR DA INTEGRAÇÃO DE ASSOCIAÇÃO GENÔMICA AMPLA E REDES DE COEXPRESSÃO GÊNICA

Dissertação apresentada ao Centro de Biociências e Biotecnologia da Universidade Estadual do Norte Fluminense Darcy Ribeiro, como parte das exigências para obtenção do título de Mestre em Biotecnologia Vegetal.

Aprovado em 17 de janeiro de 2022

Comissão examinadora:

Dr. Luiz Eduardo Del Bem – UFMG

Rodrigo Nunes da Fonseca

Dr. Rodrigo Nunes da Fonseca – UFRJ

Pinto

Dr. Vitor Batista Pinto – UENF

Dr. Thiago Motta Venancio – UENF

(Orientador)

AGRADECIMENTOS

Aos meus pais, Wanda e Marcelo, que sempre me incentivaram a buscar a excelência nos estudos e contribuíram grandemente para a formação dos meus princípios e valores.

Aos meus maiores tesouros, minha esposa Lara e minha filha Elis, que me motivam a ser melhor a cada dia, e fazem cada vitória valer a pena e cada dificuldade ser mais leve.

Ao meu orientador, Thiago Venancio, por ser um grande amigo e um exemplo de cientista e orientador, e pela formação científica de tanto tempo que me abriu tantas portas.

Aos meus colegas de laboratório – Hemanoel Passarelli, Francisnei Pedrosa, Dayana Turquetti, Cláudio Benício, Sarah Henaut e Isabella Oliveira – pelas valiosas discussões em reuniões de laboratório e pelo convívio, ainda que virtual, nesses dois últimos anos. Em especial, agradeço ao Francisnei Pedrosa por sempre resolver os problemas técnicos no laboratório, como falta de internet e problemas nas máquinas, e ao Hemanoel Passarelli por me ajudar com dicas valiosas para a seleção para o doutorado no exterior.

À Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF) por ter sido minha segunda casa (e, às vezes, primeira) por 6 anos.

Aos docentes e discentes do Programa de Pós-Graduação em Biotecnologia Vegetal por contribuírem para a minha formação formação pessoal e professional. Agradeço especialmente à secretária Margareth, um exemplo de profissionalismo, sempre disposta a ajudar os estudantes como puder.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

BIOGRAFIA

Fabrício de Almeida Silva, filho de Marcelo da Silva e Wanda Lúcia Campos de Almeida, nasceu em Campos dos Goytacazes, Rio de Janeiro, Brasil, no dia 07 de julho de 1997.

Em março de 2016, iniciou o curso de graduação em Ciências Biológicas (habilitação licenciatura) na Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), Campos dos Goytacazes, Rio de Janeiro, Brasil e concluiu a graduação em dezembro de 2019.

Em março de 2020, ingressou no Programa de Pós-Graduação em Biotecnologia Vegetal na Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), Campos dos Goytacazes, Rio de Janeiro, Brasil. De março de 2020 a novembro de 2021, desenvolveu sua pesquisa no Programa de Pós-Graduação em Biotecnologia Vegetal, sob orientação do Prof. Dr. Thiago Motta Venancio.

SUMÁRIO

RESUMO

Almeida-Silva, Fabrício, M.Sc., Universidade Estadual do Norte Fluminense Darcy Ribeiro, janeiro de 2022. **Identificação e priorização de genes de resistência a estresses bióticos em soja (*Glycine max* L. Merr.) a partir da integração de associação genômica ampla e redes de coexpressão gênica**. Orientador: Thiago Motta Venancio.

Ao longo dos últimos anos, estudos de associação genômica ampla identificaram diversos marcadores moleculares associados à resistência a estresses bióticos. Entretanto, a identificação de genes causais a partir dos marcadores ainda é um desafio. A integração de dados genéticos com dados transcriptômicos tem se tornado uma alternativa promissora para solucionar esse problema. Nesse sentido, os dois primeiros capítulos desta dissertação dedicam-se à apresentação de novos *softwares* desenvolvidos para identificar genes causais de alta confiança: i. *BioNERO*, um pacote R destinado a inferir redes regulatórias e de coexpressão a partir de dados transcriptômicos e; ii. *cageminer*, um pacote R destinado a integrar redes de coexpressão e marcadores moleculares para identificar e priorizar genes candidatos associados a características quantitativas. No terceiro e quarto capítulo, aplicamos os *softwares* desenvolvidos para identificar e priorizar genes de soja (*Glycine max*) envolvidos na resistência a fungos fitopatogênicos e pragas, respectivamente. No terceiro capítulo, identificamos 188, 56, 11, 8, e 3 genes candidatos de alta confiança para resistência a *Fusarium virguliforme, F. graminearum, Cadophora gregata, Macrophomina phaseolina* e *Phakopsora pachyrhizi*, respectivamente. No quarto capítulo, identificamos 171, 7, e 228 genes candidatos de alta confiança para resistência a *Aphis glycines, Spodoptera litura,* e *Heterodera glycines*, respectivamente. Os genes candidatos priorizados estão altamente conservados no pangenoma da soja cultivada e, de modo geral, desempenham papel em processos relacionados à imunidade, como sinalização, estresse oxidativo, reconhecimento de padrões e formação de barreiras físicas. Ainda, identificamos os acessos mais resistentes do banco de germoplasma do USDA com base no número de alelos de resistência a cada patógeno. Os acessos mais resistentes não atingem o potencial máximo, indicando que há espaço para piramidar alelos de resistência em programas de melhoramento ou por meio de edição genômica.

Palavras-chave: RNA-seq, QTL, genômica populacional, biotecnologia.

ABSTRACT

Almeida-Silva, Fabricio, M.Sc., Universidade Estadual do Norte Fluminense Darcy Ribeiro, January 2022. **Identification and prioritization of soybean (*Glycine max* L. Merr.) resistance genes against biotic stresses by integrating genome-wide association studies and gene coexpression networks.** Advisor: Thiago Motta Venancio.

Over the last years, genome-wide association studies have identified molecular markers associated with resistance to biotic stresses. However, identifying causative genes from markers remains a challenge. Integrating genetic data with transcriptome data has become a promising alternative to address this problem. In this sense, the first two chapters of this dissertation describe novel softwares we have developed to identify high-confidence causative genes: i. *BioNERO*, an R package to infer regulatory and coexpression networks from transcriptome data and; ii. *cageminer*, an R package that integrates coexpression networks and molecular markers to identify and prioritize candidate genes associated with quantitative traits. In the third and fourth chapters, we applied these softwares to identify and prioritize soybean (*Glycine max*) genes involved in resistance to phytopathogenic fungi and pests. In the third chapter, we identified 188, 56, 11, 8, and 3 high-confidence resistance genes against *Fusarium virguliforme, F. graminearum, Cadophora gregata, Macrophomina phaseolina,* and *Phakopsora pachyrhizi*, respectively. In the fourth chapter, we identified 171, 7, and 228 high-confidence candidate resistance genes against *A. glycines, S. litura,* and *H. glycines*, respectively. Overall, the prioritized candidates are highly conserved in the pangenome of cultivated soybeans and play a role in immunity-related processes, such as signaling, oxidative stress, pattern recognition, and formation of physical barriers. Further, we identified the most resistant accessions against each pathogen in the USDA germplasm based on the number of resistance alleles. The most resistant accessions do not reach the maximum potential, indicating that there is room for allele pyramiding in breeding programs or through gene editing.

Keywords: RNA-seq, QTL, population genomics, biotechnology.

Introdução Geral

A cultura da soja

A soja (*Glycine max* (L.) Merr.) é a principal leguminosa produzida no mundo. O genoma da soja, publicado há uma década, apresenta fortes assinaturas de dois eventos de poliploidização, que ocorreram há cerca de 58 e 13 milhões de anos, respectivamente (Schmutz et al., 2010; Severin et al., 2010). Em consequência desses eventos, 75% dos genes da soja estão presentes em múltiplas cópias, representando uma característica distintiva dessa espécie em relação às demais de sua família (Schmutz et al., 2010).

Originada e domesticada no leste da China há cerca de 6-9 mil anos (Sedivy et al., 2017), a cultura da soja foi introduzida no Brasil no estado da Bahia e, posteriormente, no Rio Grande do Sul. A partir da década de 1960, o melhoramento genético da soja permitiu a transferência da maior fração da cultura para a região Centro-Oeste, onde havia vastas áreas cultiváveis, acelerando a produção dessa *commodity* agrícola (Cattelan and Dall'Agnol, 2018). Atualmente, o estado do Mato Grosso é o maior produtor nacional de soja, seguido do Paraná e Rio Grande do Sul (Cattelan and Dall'Agnol, 2018).

Atualmente, o Brasil é o maior produtor de soja do mundo, seguido pelos Estados Unidos. O lucro obtido com exportações de soja (grão, farelo e óleo) na safra 2020/2021 foi de 35,23 bilhões de dólares, o que corresponde a 1,13% do PIB do ano (EMBRAPA SOJA, 2019). Além da relevância econômica nacional e global, os grãos representam uma fração significativa da dieta humana e animal (Yang et al., 2019). A soja representa direta e indiretamente 70% das fontes de proteínas da dieta e 30% das fontes de óleo, enfatizando sua relevância para a segurança alimentar global (Gao et al., 2018).

Estresse biótico em soja

O estresse biótico, definido como o estresse causado por organismos vivos (i.e., patógenos e pragas), é um dos principais desafios para a produção global de soja (Kankanala et al., 2019). As doenças de soja causam um prejuízo anual de 4,55 bilhões de dólares para os EUA (Bandara et al., 2020). As pragas, representadas majoritariamente por insetos, causam prejuízo econômico de 17,7 bilhões de dólares no Brasil (Oliveira et al., 2014).

As principais doenças fúngicas em soja incluem ferrugem, míldio, síndrome da morte súbita e antracnose (Kankanala et al., 2019). As doenças bacterianas causam danos como necrose, nanismo e lesão foliar (Wille et al., 2019). As doenças virais são causadas, principalmente, pelos vírus do mosaico da soja, vírus do mosqueado do feijão e vírus do mosqueado do amendoim, que causam diversos sintomas distintos (Chang et al., 2016). Ainda, a soja pode ser parasitada por nematoides e oomicetos, devastando grandes áreas de plantação (Rubiales et al., 2015).

As plantas evoluíram diversos mecanismos de defesa contra o ataque de patógenos e pragas. O estresse biótico ativa uma cascata de sinalização decorrente do reconhecimento de padrões moleculares associados a patógenos (Kankanala et al., 2019). Esses mecanismos de resposta geram alterações na expressão gênica da planta, podendo inibir ou estimular a expressão de certos genes (Cohen and Leach, 2019). Ao nível transcricional, a resposta a patógenos biotróficos ativa genes de vias dependentes de ácido salicílico, enquanto patógenos hemibiotróficos e necrotróficos ativam genes de sinalização por etileno e ácido jasmônico (Glazebrook, 2005).

Estudos de associação genômica ampla (GWAS)

O mapeamento de *loci* de caracteres quantitativos (QTL, do inglês *Quantitative Trait Loci*) a partir de cruzamentos biparentais é uma técnica amplamente utilizada (da Silva et al., 2019; de Ronne et al., 2020; Guo et al., 2020a; Hackenberg et al., 2020). Entretanto, essa técnica apresenta limitações, pois permite descrever somente a diversidade alélica limitada das linhagens parentais, além de apresentar baixa resolução genômica (Vuong et al., 2015). Os estudos de associação genômica ampla (GWAS, do inglês *Genome-Wide Association Studies*) representam um método alternativo para associar variantes genéticas a características de interesse (Peat et al., 2020). Os GWAS apresentam maior poder estatístico e maior resolução genômica, pois possibilitam análises com populações grandes e geneticamente diversas.

Diversos GWAS já foram conduzidos para identificar polimorfismos de nucleotídeo único (SNPs, do inglês *Single Nucleotide Polymorphisms*) associados a resistência a estresse biótico em soja (Boudhrioua et al., 2020; Maldonado Dos Santos et al., 2019; Rolling et al., 2020; Swaminathan et al., 2019; Vinholes et al., 2019). Por exemplo, Chang e colaboradores identificaram *loci* associados a resistência a 11 doenças de soja (Chang et al., 2016) e diversos insetos-praga (Chang and Hartman, 2017). Zhang e colaboradores identificaram no acesso PI 82278 o maior número de alelos de resistência à síndrome da morte súbita, uma das doenças fúngicas mais severas em soja (Zhang et al., 2015). Acessos com grande número de alelos de resistência podem ser usados para iniciar programas de melhoramento genético e seleção assistida por marcadores, gerando populações mais resistentes ao estresse biótico.

A identificação de genes candidatos a partir dos SNPs significativos obtidos pelos GWAS ainda é arbitrária. Diversos autores selecionam como candidatos os genes em alto desequilíbrio de ligação com os SNPs significativos (Boudhrioua et al., 2020; Maldonado Dos Santos et al., 2019; Tran et al., 2019). Outros, por sua vez, selecionam os genes localizados em intervalos arbitrários (*e.g.,* 50 kbp) em relação aos SNPs significativos (Moellers et al., 2017; Zhao et al., 2017). Ambos os critérios apresentam taxas altas de falso-positivos e falso-negativos, seja por uma quantidade variável de recombinação no painel de associação ou pela seleção de intervalos que desconsiderem genes causais (Michno et al., 2020).

Biologia de sistemas e redes de coexpressão

Os recentes avanços nas tecnologias de sequenciamento possibilitaram o surgimento e desenvolvimento da biologia de sistemas (Lavarenne et al., 2018). A biologia de sistemas consiste na análise de componentes moleculares (*e.g.,* genes, proteínas, metabólitos), não como entidades independentes, mas como partes de uma rede complexa e dinâmica (Gaudinier and Brady, 2016). As redes de coexpressão (GCNs, do inglês *Gene Coexpression Networks*) são representadas por grafos cujos vértices representam genes, e arestas representam as correlações entre pares de genes (Fuller et al., 2007).

As GCNs têm sido amplamente utilizadas para estudar a regulação transcricional e evolução de plantas (Du et al., 2017; Huang et al., 2019; Wisecaver et al., 2017). Por exemplo, Wu e colaboradores reconstruíram uma GCN em soja e identificaram um grupo de genes relacionados à nodulação em leguminosas (Wu et al., 2019). Almeida-Silva e colaboradores reconstruíram uma GCN com 1284 amostras de RNA-seq de soja e identificaram potenciais reguladores de vias como biossíntese de lignina, resposta a fungos e fotossíntese, além de elucidar a dinâmica da regulação transcricional de genes duplicados (Almeida-Silva et al., 2020).

Devido ao seu potencial de detectar padrões em larga escala, as GCNs têm sido integradas a métodos de genômica populacional, como GWAS (Guo et al., 2020b; Schaefer et al., 2018). A base lógica dessa abordagem deriva da pressuposição de que genes pertencentes a um mesmo processo biológico são co-regulados (*i.e.*, coexpressos). Usando essa abordagem integrativa, Schaefer e colaboradores identificaram genes em milho (*Zea mays*) relacionados ao acúmulo de diversos íons (Guo et al., 2020b; Schaefer et al., 2018). Essa abordagem promissora pode acelerar e otimizar a identificação de genes candidatos envolvidos em processos biológicos de interesse.

Esta dissertação buscou identificar genes candidatos de alta confiança para resistência a estresses bióticos em soja. Os capítulos seguintes são compilações de artigos independentes produzidos durante o curso de mestrado. Ao final da dissertação, apresentamos título e resumo de outros artigos produzidos durante o curso de mestrado, mas que não estão vinculados a esse projeto principal.

Referências

Almeida-Silva, F., Moharana, K.C., Machado, F.B., and Venancio, T.M. (2020). Exploring the complexity of soybean (Glycine max) transcriptional regulation using global gene co-expression networks. Planta **252**: 1–12.

Bandara, A.Y., Weerasooriya, D.K., Bradley, C.A., Allen, T.W., and Esker, P.D. (2020). Dissecting the economic impact of soybean diseases in the United States over two decades. PLoS One **15**: 1–28.

Boudhrioua, C., Bastien, M., Torkamaneh, D., and Belzile, F. (2020). Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. BMC Plant Biol. **20**: 195.

Cattelan, A.J. and Dall'Agnol, A. (2018). The rapid soybean growth in Brazil. OCL - Oilseeds fats, Crop. Lipids **25**: 1–12.

Chang, H.X. and Hartman, G.L. (2017). Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Front. Plant Sci. **8**: 1–12.

Chang, H.X., Lipka, A.E., Domier, L.L., and Hartman, G.L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Phytopathology **106**: 1139–1151.

Cohen, S.P. and Leach, J.E. (2019). Abiotic and biotic stresses induce a core transcriptome response in rice. Sci. Rep. **9**: 1–11.

Du, J., Wang, S., He, C., Zhou, B., Ruan, Y.-L., and Shou, H. (2017). Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. J. Exp. Bot. **68**: erw460.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm. Genome **18**: 463–472.

Gao, H., Wang, Y., Li, W., Gu, Y., Lai, Y., Bi, Y., and He, C. (2018). Transcriptomic comparison reveals genetic variation potentially underlying seed developmental evolution of soybeans. J. Exp. Bot. **69**: 5089–5104.

Gaudinier, A. and Brady, S.M. (2016). Mapping Transcriptional Networks in Plants: Data-Driven Discovery of Novel Biological Mechanisms. Annu. Rev. Plant Biol. **67**: 575–594.

Glazebrook, J. (2005). Contrasting Mechanisms of Defense Against Biotrophic and

Necrotrophic Pathogens. Annu. Rev. Phytopathol. **43**: 205–227.

Guo, J., Li, C., Zhang, X., Li, Y., Zhang, D., Shi, Y., Song, Y., Li, Y., Yang, D., and Wang, T. (2020a). Transcriptome and GWAS analyses reveal candidate gene for seminal root length of maize seedlings under drought stress. Plant Sci. **292**.

Guo, W. et al. (2020b). Characterization of Pingliang xiaoheidou (ZDD 11047), a soybean variety with resistance to soybean cyst nematode Heterodera glycines. Plant Mol. Biol. **103**: 253–267.

Hackenberg, D. et al. (2020). Identification and QTL mapping of resistance to Turnip yellows virus (TuYV) in oilseed rape, Brassica napus. Theor. Appl. Genet. **133**: 383–393.

Huang, A.C., Jiang, T., Liu, Y.X., Bai, Y.C., Reed, J., Qu, B., Goossens, A., Nützmann, H.W., Bai, Y., and Osbourn, A. (2019). A specialized metabolic network selectively modulates Arabidopsis root microbiota. Science (80-.). **364**.

Kankanala, P., Nandety, R.S., and Mysore, K.S. (2019). Genomics of Plant Disease Resistance in Legumes. Front. Plant Sci. **10**: 1–20.

Lavarenne, J., Guyomarc'h, S., Sallaud, C., Gantet, P., and Lucas, M. (2018). The Spring of Systems Biology-Driven Breeding. Trends Plant Sci. **23**: 706–720.

Maldonado Dos Santos, J.V., Ferreira, E.G.C., Passianotto, A.L.D.L., Brumer, B.B., Santos, A.B. Dos, Soares, R.M., Torkamaneh, D., Arias, C.A.A., Belzile, F., Abdelnoor, R.V., and Marcelino-Guimarães, F.C. (2019). Association mapping of a locus that confers southern stem canker resistance in soybean and SNP marker development. BMC Genomics **20**: 1–13.

Michno, J.M., Liu, J., Jeffers, J.R., Stupar, R.M., and Myers, C.L. (2020). Identification of nodulation-related genes in Medicago truncatula using genome-wide association studies and co-expression networks. Plant Direct **4**: 1–10.

Moellers, T.C., Singh, A., Zhang, J., Brungardt, J., Kabbage, M., Mueller, D.S., Grau, C.R., Ranjan, A., Smith, D.L., Chowda-Reddy, R. V., and Singh, A.K. (2017). Main and epistatic loci studies in soybean for Sclerotinia sclerotiorum resistance reveal multiple modes of resistance in multi-environments. Sci. Rep. **7**: 1–13.

Oliveira, C.M., Auad, A.M., Mendes, S.M., and Frizzas, M.R. (2014). Crop losses and the economic impact of insect pests on Brazilian agriculture. Crop Prot. **56**: 50–54.

Peat, G., Jones, W., Nuhn, M., Marugán, J.C., Newell, W., Dunham, I., and Zerbino,

D. (2020). The open targets post-GWAS analysis pipeline. Bioinformatics: 1–2.

Rolling, W., Lake, R., Dorrance, A.E., and McHale, L.K. (2020). Genome-wide association analyses of quantitative disease resistance in diverse sets of soybean [Glycine max (L.) Merr.] plant introductions. PLoS One **15**: 1–28.

de Ronne, M., Labbé, C., Lebreton, A., Sonah, H., Deshmukh, R., Jean, M., Belzile, F., O'Donoughue, L., and Bélanger, R. (2020). Integrated QTL mapping, gene expression and nucleotide variation analyses to investigate complex quantitative traits: a case study with the soybean–Phytophthora sojae interaction. Plant Biotechnol. J. **18**: 1492–1494.

Rubiales, D., Fondevilla, S., Chen, W., Gentzbittel, L., Higgins, T.J.V., Castillejo, M.A., Singh, K.B., and Rispail, N. (2015). Achievements and Challenges in Legume Breeding for Pest and Disease Resistance. CRC. Crit. Rev. Plant Sci. **34**: 195–236.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell **30**: 2922–2942.

Schmutz, J. et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183.

Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. **214**: 539–553.

Severin, A.J. et al. (2010). RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biol. **10**: 160.

da Silva, M.P., Klepadlo, M., Gbur, E.E., Pereira, A., Mason, R.E., Rupe, J.C., Bluhm, B.H., Wood, L., Mozzoni, L.A., and Chen, P. (2019). QTL mapping of charcoal rot resistance in PI 567562A soybean accession. Crop Sci. **59**: 474–479.

Swaminathan, S., Das, A., Assefa, T., Knight, J.M., Da Silva, A.F., Carvalho, J.P.S., Hartman, G.L., Huang, X., Leandro, L.F., Cianzio, S.R., and Bhattacharyya, M.K. (2019). Genome wide association study identifies novel single nucleotide polymorphic loci and candidate genes involved in soybean sudden death syndrome resistance. PLoS One **14**: 1–21.

Tran, D.T., Steketee, C.J., Boehm, J.D., Noe, J., and Li, Z. (2019). Genome-wide association analysis pinpoints additional major genomic regions conferring resistance to soybean cyst nematode (Heterodera glycines ichinohe). Front. Plant Sci. **10**: 1–13.

Vinholes, P., Rosado, R., Roberts, P., Borém, A., and Schuster, I. (2019). Single nucleotide polymorphism-based haplotypes associated with charcoal rot resistance in Brazilian soybean germplasm. Agron. J. **111**: 182–192.

Vuong, T.D., Sonah, H., Meinhardt, C.G., Deshmukh, R., Kadam, S., Nelson, R.L., Shannon, J.G., and Nguyen, H.T. (2015). Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC Genomics **16**: 1–13.

Wille, L., Messmer, M.M., Studer, B., and Hohmann, P. (2019). Insights to plant–microbe interactions provide opportunities to improve resistance breeding against root diseases in grain legumes. Plant Cell Environ. **42**: 20–40.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A Global Co-expression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell: tpc.00009.2017.

Wu, Z., Wang, M., Yang, S., Chen, S., Chen, X., Liu, C., Wang, S., Wang, H., Zhang, B., Liu, H., Qin, R., and Wang, X. (2019). A global coexpression network of soybean genes gives insights into the evolution of nodulation in nonlegumes and legumes. New Phytol. **223**: 2104–2119.

Yang, S., Miao, L., He, J., Zhang, K., Li, Y., and Gai, J. (2019). Dynamic transcriptome changes related to oil accumulation in developing soybean seeds. Int. J. Mol. Sci. **20**.

Zhang, J., Singh, A., Mueller, D.S., and Singh, A.K. (2015). Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. Plant J. **84**: 1124–1136.

Zhao, X., Teng, W., Li, Y., Liu, D., Cao, G., Li, D., Qiu, L., Zheng, H., Han, Y., and Li, W. (2017). Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5.7. BMC Genomics **18**: 1–10.

CHAPTER 1:

BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction

Chapter 1: BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction

Fabricio Almeida-Silva[1*] and Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Short communication

Situation: Published (DOI: 10.1007/s10142-021-00821-9)

Journal: Functional and Integrative Genomics

Impact factor: 3.410; CiteScore: 5.9; Highest percentile: 68%

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

ABSTRACT

Summary

Currently, standard network analysis workflows rely on many different packages, often requiring users to have a solid statistics and programming background. Here, we present BioNERO, an R package that aims to integrate all aspects of network analysis workflows, including expression data preprocessing, gene coexpression and regulatory network inference, functional analyses, and intra and interspecies network comparisons. The state-of-the-art methods implemented in BioNERO ensure that users can perform all analyses with a single package in a simple pipeline, without needing to learn a myriad of package-specific syntaxes. BioNERO offers a user-friendly framework that can be easily incorporated in systems biology pipelines.

Availability and implementation

The package is available at Bioconductor (http://bioconductor.org/packages/BioNERO).

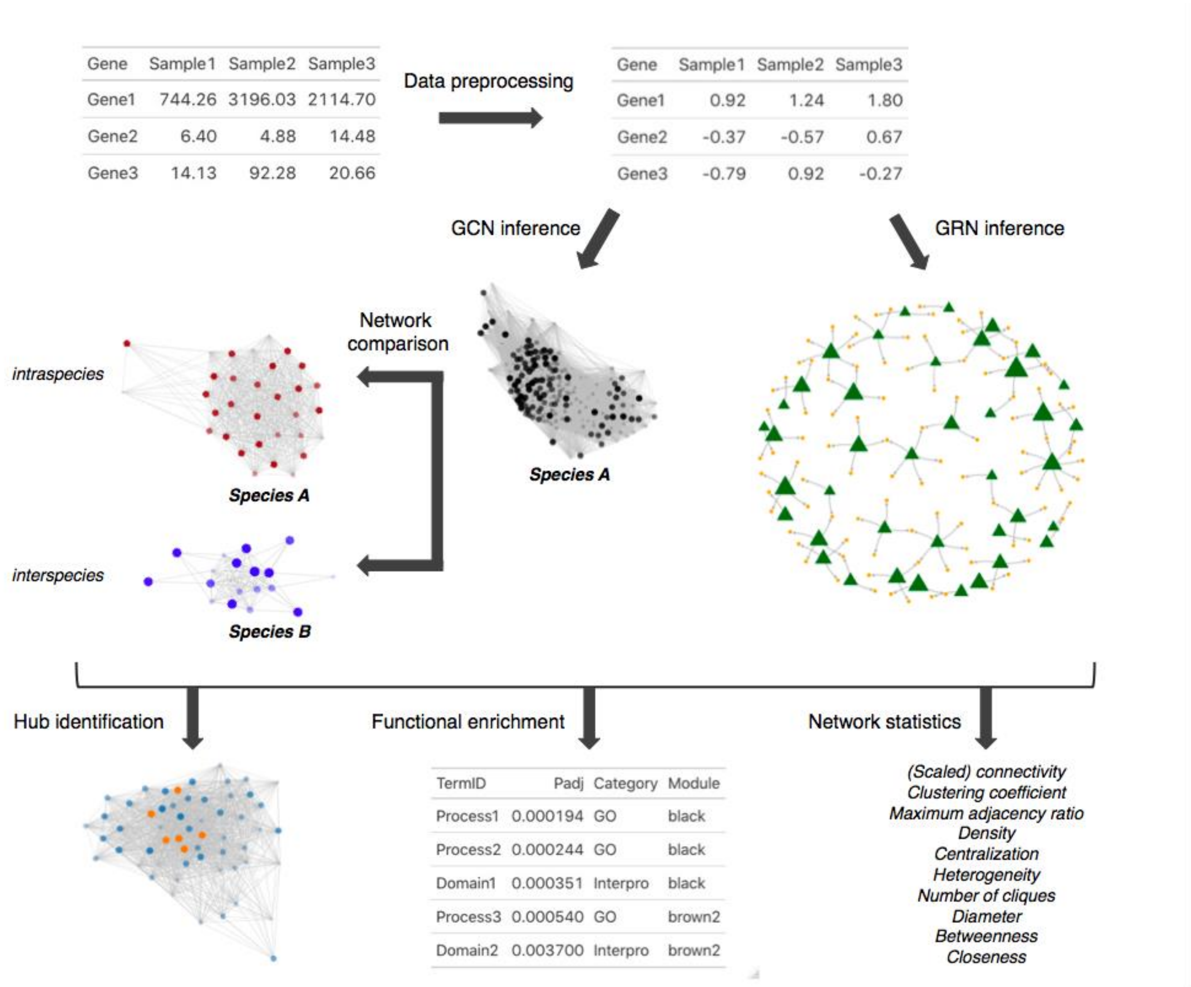
1 Introduction

To date, several packages have been developed to infer gene coexpression networks (GCNs) and gene regulatory networks (GRN) from expression data, such as WGCNA (Langfelder and Horvath, 2008), CEMiTool (Russo et al., 2018), petal (Petereit et al., 2016), and minet (Meyer et al., 2008). However, none of them can handle all aspects of network analysis workflows, and users are required to use other packages to build a standard analysis pipeline. Further, network inference requires a solid linear algebra and statistics background, resulting in a struggle for inexperienced researchers to properly preprocess their expression data and extract biologically meaningful information from the inferred networks.

Here, we present *BioNERO* (Biological Network Reconstruction Omnibus), an R/Bioconductor package that integrates all steps of network inference workflows in a single package. *BioNERO* uses state-of-the-art methods to preprocess expression data, infer GCNs and GRNs from expression data, analyze networks for biological interpretations, and compare networks within and across species. Additionally, *BioNERO* can be used to explore topological properties of protein-protein interaction networks, such as hub identification and community detection.

2 Methods

BioNERO is an R package that integrates existing functionalities and introduces new ones. The input data can be common Bioconductor classes, such as SummarizedExperiment objects (Morgan et al., 2020) for expression data, or basic R object classes, ensuring interoperability with other packages. Long-running functions, such as that used for Fisher’s exact tests in overrepresentation analyses, have been parallelized with BiocParallel (Morgan et al., 2021) to increase speed. A summary of the BioNERO algorithm is described in Figure 1.



Gene	Sample1	Sample2	Sample3
Gene1	744.26	3196.03	2114.70
Gene2	6.40	4.88	14.48
Gene3	14.13	92.28	20.66

Gene	Sample1	Sample2	Sample3
Gene1	0.92	1.24	1.80
Gene2	-0.37	-0.57	0.67
Gene3	-0.79	0.92	-0.27

TermID	Padj	Category	Module
Process1	0.000194	GO	black
Process2	0.000244	GO	black
Domain1	0.000351	Interpro	black
Process3	0.000540	GO	brown2
Domain2	0.003700	Interpro	brown2

Figure 1. Summary of the BioNERO algorithm. From a raw gene expression matrix, users can preprocess their data and infer GRNs and GCNs, and the latter can be used for comparisons within and across species. The possible downstream network analyses are hub identification, functional enrichment, gene-/module-trait associations, visualization, community detection (for GRNs and PPI), and calculation of network statistics.

3 Results

3.1 Data preprocessing

Networks inferred from unfiltered data often do not satisfy the scale-free topology (SFT) assumption. Although this can be a property of the input data (particularly for heterogenous data sets), this issue mainly results from a lack of systematic preprocessing. With *BioNERO*, users can preprocess their expression data prior to network inference to i. remove missing data; ii. remove genes with low expression across samples; iii. remove outliers; iv. select genes with the highest variances (optional) and; v. remove confounders that could introduce false-positive correlations.

Outlier removal is based on the standardized connectivity (Zk) method, which can detect outliers that other methods (*e.g.,* hierarchical clustering) cannot and, hence, it is more suitable for network analysis (Oldham et al., 2012). Adjusting for confounders relies on a principal component-based method implemented in the Bioconductor package sva (Parsana et al., 2019; Leek et al., 2021). The resulting expression matrix can be quantile normalized to make it suitable for parametric tests. Count data can also be variance stabilizing transformed with DESeq2's algorithm (Love et al., 2014) to make the expression matrix approximately homoscedastic.

3.2 Gene coexpression network inference

GCN inference from expression data in BioNERO relies on the popular Weighted Gene Coexpression Network Analysis (WGCNA) algorithm, implemented in the WGCNA R package (Langfelder and Horvath, 2008). A matrix of pairwise gene-gene correlations can be calculated with Pearson's r, Spearman's ρ, or biweight midcorrelation (median-based, which is less sensible to outliers), and it is further transformed to an adjacency matrix to amplify disparities. Users can infer three types of GCNs (signed, signed hybrid, or unsigned), and network type affects the way adjacency matrices are calculated. Signed networks (default) preserve correlation signs, so positive and negative correlation coefficients are interpreted as different. Signed hybrid networks treat all negative correlation coefficients as zero, so only positive correlations are considered. Unsigned networks ignore correlation signs, so positive and negative values are not distinguished.

3.3 Gene regulatory network inference

Different GRN inference algorithms can be the best performers depending on the benchmark expression data set, as demonstrated by Marbach *et al.* (2012). This observation inspired the "wisdom of the crowds" principle for GRN inference, which consists in calculating average ranks for all edges across different algorithms to obtain consensus, high-confidence edges (Marbach et al., 2012). BioNERO offers three widely used GRN inference algorithms: GENIE3, imported from the R package GENIE3 (Huynh-Thu et al., 2010); ARACNE, imported from the R package minet (Margolin et al., 2006), and CLR, also from minet (Faith et al., 2007). However, choosing the most appropriate number of top edges to keep is a persisting bottleneck, and users often pick an arbitrary number. We implemented a method to simulate different networks by splitting the graph in *n* subgraphs, each containing the top *n*th quantiles. Then, we calculate SFT fit statistics for each subgraph and select the top number of edges that leads to the best SFT fit.

3.4 Module detection and network statistics

Module detection in GCNs relies on the dynamicTreeCut (Langfelder et al., 2008) package as implemented in WGCNA. After module detection, very similar modules can be merged if the correlation of their eigengenes (first principal component) is greater than a given threshold (by default, 0.8). Module stability can be assessed to test if the network topology depends on a small subset of samples. For physical networks (GRNs and PPI), community detection relies on the *cluster_()* functions from the R package igraph (Csardi and Nepusz, 2006), and several methods are available, such as infomap (default), edge betweenness, fast greedy, label propagation, walktrap, and louvain. Additionally, *BioNERO* imports igraph to calculate main network statistics, namely connectivity, scaled connectivity, clustering coefficient, maximum adjacency ratio, density, centralization, heterogeneity, number of cliques, diameter, betweenness, and closeness.

3.5 Functional analyses and network exploration

After inferring GCNs, users can input a data frame of gene annotation to perform module overrepresentation analysis (ORA) and test if modules are enriched in genes associated with a particular biological process, metabolic pathway, protein domain, or any other annotation. ORA can also be performed for a user-defined gene set, even if they are in different modules. ORA results can be interpreted in combination with gene-/module-trait associations, which identify genes and/or modules whose expression levels increase or decrease in a particular condition. Further, users can identify network hubs as the top 10% most highly connected nodes (PPI and GRNs) or as the intersection between the top 10% most highly connected genes and genes with module membership ≥0.8 (GCN), as defined in a previous work from our group (Almeida-Silva et al., 2020). Users can also extract subgraphs for a particular module or custom gene set, and they can be used for visualization or calculation of statistics. For all subgraph extractions, users can verify if the graphs fit the SFT, a characteristic of real-world biological networks (Barabási et al., 2011).

3.6 Exploratory analyses and data visualization

BioNERO offers data visualization functions for exploratory analyses (principal component analysis and heatmaps) and visualization of results (Figure 2). The graphical functions for gene-/module-trait associations, dendrogram of genes and modules, and eigengene network rely on the base plotting system, with functions imported from WGCNA (Langfelder and Horvath, 2008). Gene expression and sample correlation heatmaps rely on the ComplexHeatmap package (Gu et al., 2016). The ggplot2 system is used for all other data visualizations, namely principal component analysis plots, module expression profile, frequency of genes per module, and network plots. Static network plots rely on the ggnetwork package (Briatte, 2021), while interactive networks are powered by the D3 Javascript library with the R package networkD3 (Allaire et al., 2017).

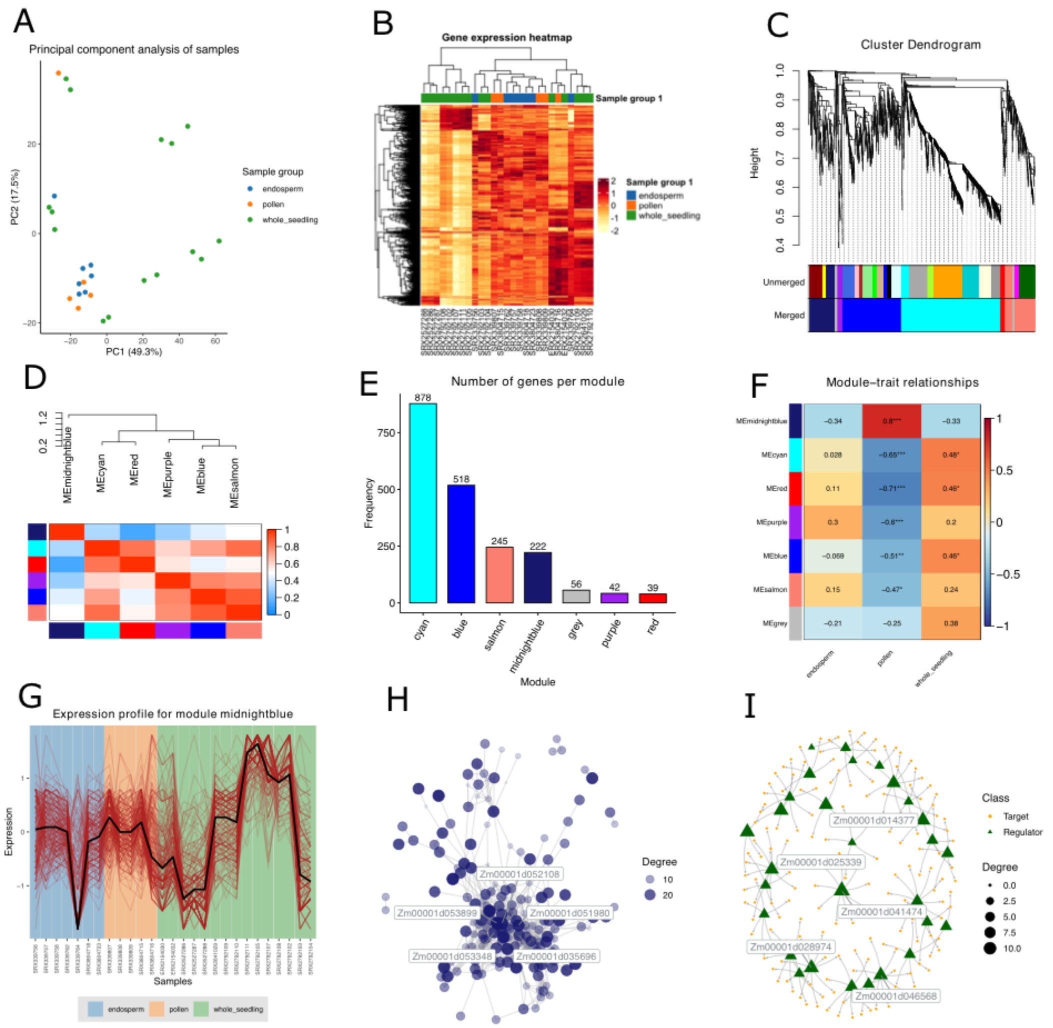


Figure 2. Overview of plots that can be created with BioNERO's built-in functions. A. Principal component analysis of samples. Users can plot principal component (PC) 1 vs PC2, PC1 vs PC3, and PC2 vs PC3. Variance explained by each PC is included in the axis labels. B. Gene expression heatmap. Genes and samples can be hierarchically clustered. Gene and sample metadata can be given as input, so a color code will be added to rows/columns. C. Dendrogram of genes and modules before and after merging similar modules. D. Eigengene network. Colors display correlations between module eigengenes. E. Absolute frequency of genes per module. F. Module-trait correlations. This heatmap shows modules comprising genes whose expression levels significantly increase or decrease in a particular condition. G. Gene expression profile across samples for a particular module. H. Static GCN visualization. Labels indicate hub genes. I. Static GRN visualization. Green triangles represent regulators, while gold circles represent targets. All plots were created with the example data from the package's vignettes.

3.7 Network comparison

GCNs inferred from different expression sets have similarities and divergences. BioNERO offers two network comparison approaches, namely consensus module identification and module preservation. Consensus modules are gene modules that co-occur in networks inferred from independent expression sets, and they can be used to explore core components of the studied phenotype that are not affected by experimental effects or natural biological variation. While consensus modules identification focuses on the similarities between networks, module preservation focuses on the differences, and it can be used to explore patterns of transcriptional divergence within and across species. Consensus module identification relies on the R package WGCNA, while network preservation relies on non-parametric permutation tests implemented in the R package NetRep (Ritchie et al., 2016). For interspecies comparisons, *BioNERO* can interoperate with OrthoFinder (Emms and Kelly, 2015) to analyze expression profiles at the orthogroup level.

3.8 Comparing BioNERO to other packages

We compared BioNERO to the main network inference-related R packages, namely WGCNA (Langfelder and Horvath, 2008), CEMiTool (Russo et al., 2018), petal (Petereit et al., 2016), minet (Meyer et al., 2008), and GENIE3 (Huynh-Thu et al., 2010). All packages were given points based on the functionalities they offer (Table 1). BioNERO outperforms all existing network inference-related R packages, as it acts as a hub by integrating different functionalities. Although some functionalities available in BioNERO are already included in other packages, none of them include all of BioNERO's functionalities. The second package in number of functionalities is WGCNA, with only half of BioNERO's potential. Although CEMiTool is easy to use and can infer GCNs with a single function, it has fewer functionalities than WGCNA. The other R packages are limited to a specific goal and, hence, they have the fewest points.

Table 1. Comparative view of functionalities in BioNERO and other network inference-related R packages.

Functionalities	BioNERO	WGCNA	CEMiTool	petal	minet	GENIE3
Gene filtering	1	0	1	0	0	0
Correction for confounders	1	0	0	0	0	0
GCN inference (signed, signed hybrid, unsigned)	3	3	2	1	1	0
GRN inference (MI, RF, PC)	3	0	0	0	2	1
Module functional enrichment	1	1	1	0	0	0
Topology-based network filtering	1	0	0	1	0	0
Static network visualization	1	0	1	0	0	0
Interactive network visualization	1	0	0	0	0	0
Intraspecies network comparison	1	1	0	0	0	0
Interspecies network comparison	1	0	0	0	0	0
Calculation of network statistics	1	1	0	1	0	0
Community detection (GCN, GRN, PPI)	3	1	1	1	0	0
Hub identification	1	1	1	0	0	0
Total points	**19**	**8**	**7**	**4**	**3**	**1**

GCN: Gene Coexpression Network. GRN: Gene Regulatory Network. PPI: Protein-Protein Interaction. MI: Mutual Information. RF: Random Forests. PC: Partial Correlations.

3.9 Application to real data sets

A use case using maize (*Zea mays*) and rice (*Oryza sativa*) gene expression data obtained from Shin *et al.* (2020) is available as Supplementary Text. The maize RNA-seq data set comprises 39,604 genes and 116 samples, while the rice RNA-seq data set comprises 35,667 genes and 265 samples.

4 Conclusions

BioNERO is a novel R package that integrates all steps of network analysis pipelines, providing users with a simple framework for GCN and GRN inference from expression data. This package can be easily integrated in systems biology pipelines and will likely accelerate biological network analysis projects.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES;

Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.

Conflicts of interest: none declared.

REFERENCES

Allaire, J.J., Gandrud, C., Russell, K., and Yetman, C.J. (2017). networkD3: D3 JavaScript Network Graphs from R.

Almeida-Silva, F., Moharana, K.C., Machado, F.B., and Venancio, T.M. (2020). Exploring the complexity of soybean (Glycine max) transcriptional regulation using global gene co-expression networks. Planta **252**: 1–12.

Barabási, A.-L., Ravasz, E., and Oltvai, Z. (2011). Hierarchical Organization of Modularity in Complex Networks. Science (80-.). **297**: 46–65.

Briatte, F. (2021). ggnetwork: Geometries to Plot Networks with ggplot2.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, complex Syst. **1695**: 1–9.

Emms, D.M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. **16**: 1–14.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. **5**: 0054–0066.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One **5**: 1–10.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**: 559.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. Bioinformatics **24**: 719–720.

Leek, J.T., Johnson, W.E., Parker, H.S., Fertig, E.J., Jaffe, A.E., Zhang, Y., Storey,

J.D., and Torres, L.C. (2021). sva: Surrogate Variable Analysis.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**: 1–21.

Marbach, D. et al. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods **9**: 796–804.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**: 1–15.

Meyer, P.E., Lafitte, F., and Bontempi, G. (2008). Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics **9**: 1–10.

Morgan, M., Obenchain, V., Hester, J., and Pagès, H. (2020). SummarizedExperiment: SummarizedExperiment container.

Morgan, M., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2021). BiocParallel: Bioconductor facilities for parallel evaluation.

Oldham, M.C., Langfelder, P., and Horvath, S. (2012). Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst. Biol. **6**: 1.

Parsana, P., Ruberman, C., Jaffe, A.E., Schatz, M.C., Battle, A., and Leek, J.T. (2019). Addressing confounding artifacts in reconstruction of gene co-expression networks. Genome Biol. **20**: 94.

Petereit, J., Smith, S., Harris, F.C., and Schlauch, K.A. (2016). petal: Co-expression network modelling in R. BMC Syst. Biol. **10**: 51.

Ritchie, S.C., Watts, S., Fearnley, L.G., Holt, K.E., Abraham, G., and Inouye, M. (2016). A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets. Cell Syst. **3**: 71–82.

Russo, P.S.T. et al. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics **19**: 56.

Shin, J., Marx, H., Richards, A., Vaneechoutte, D., Jayaraman, D., Maeda, J., Chakraborty, S., Sussman, M., Coon, J., Roy, S., Vandepoele, K., and An, J. (2020). A network-based comparative framework to study conservation and divergence of proteomes in plant phylogenies. Nucleic Acids Res.: 1–23.

CHAPTER 2:

cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks

Chapter 2: cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks

Fabricio Almeida-Silva[1*] and Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Application Note

Situation: Under review

Journal: Bioinformatics

Impact factor: 6.937; CiteScore: 9.9; Highest percentile: 98%

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

ABSTRACT

Summary

Although genome-wide association studies (GWAS) identify variants associated with traits of interest, they often fail in identifying causative genes underlying a given phenotype. Integrating GWAS and gene coexpression networks can help prioritize high-confidence candidate genes, as the expression profiles of trait-associated genes can be used to mine novel candidates. Here, we present *cageminer*, the first R package to prioritize candidate genes through the integration of GWAS and coexpression networks. Genes are considered high-confidence candidates if they pass all three filtering criteria implemented in *cageminer*, namely physical proximity to SNPs, coexpression with known trait-associated genes, and significant changes in expression levels in conditions of interest. Prioritized candidates can also be scored and ranked to select targets for experimental validation. By applying *cageminer* to a real data set, we demonstrate that it can effectively prioritize candidates, leading to >99% reductions in candidate gene lists.

Availability and implementation

The package is available at Bioconductor (http://bioconductor.org/packages/cageminer).

1 Introduction

Over the years, several genome-wide association studies (GWAS) have identified single-nucleotide polymorphisms (SNPs) associated with phenotypes of interest, such as agronomic traits in crops, production traits in livestock, and complex human disorders (Boudhrioua et al., 2020; Maldonado Dos Santos et al., 2019; Wu et al., 2020; Buzanskas et al., 2014). However, finding causative genes from SNPs remains a major bottleneck (Baxter, 2020). First, most GWAS-derived SNPs are located in non-coding portions of the genome, which can be regulatory regions very far from a causative gene (Peat et al., 2020). Further, causative variants can be in strong linkage disequilibrium (LD) with non-causative ones, leading to large LD blocks with dozens of putative candidates (Michno et al., 2020).

To address this issue, integrating GWAS with the vast amounts of RNA-seq data in public repositories has become a promising solution, particularly using gene coexpression network (GCN)-based approaches (Michno et al., 2020; Yao et al., 2020; Guo et al., 2020). Currently, the only statistical framework that automates such integration is Camoco, a Python library that identifies sets of densely connected genes for a given sliding window relative to each SNP (Schaefer et al., 2018). However, as sliding windows are expanded (*e.g.,* 50 kb), Camoco loses the ability to discover candidate genes because of background noise (Michno et al., 2020). This is a major limitation, as SNPs can be up to 2 Mb away from the causative genes if they are in distal regions (Brodie et al., 2016).

Here, we present *cageminer* (candidate gene miner), the first R/Bioconductor package that integrates GCNs and GWAS-derived SNPs to prioritize candidate genes associated with traits of interest. *cageminer* uses a guide gene-based approach to discover novel candidates that are coexpressed with known trait-associated genes and are significantly induced or repressed in conditions of interest. By relying on researchers' prior knowledge, *cageminer* can identify high-confidence candidate genes even in megabase-scale genomic intervals. This package will be instrumental in helping researchers discover genes underlying important quantitative traits.

2 Implementation

cageminer is implemented as an R package, and all input and output objects belong to base R or common Bioconductor classes to ensure interoperability with other packages. Our algorithm requires three types of input data: i. SNP positions, which must be passed as GRanges or GRangesList objects (for single trait and multiple traits, respectively) (Lawrence et al., 2013); ii. guide genes, either as a character vector or a data frame; and iii. gene coexpression network, which must be passed as a list as returned by the function *exp2gcn()* from the Bioconductor package *BioNERO* (Almeida-Silva and Venancio, 2021).

2.1 Algorithm description

cageminer identifies high-confidence candidate genes in three sequential steps (Fig. 1). In the first step, all genes within a sliding window (default: 2 Mb) relative to each SNP are selected as putative candidates. The default 2 Mb sliding window aims to minimize false-negative rates, as SNPs can be located in distal regions (Brodie et al., 2016). If the 2 Mb window returns too many genes to start with, users can simulate different window sizes and visualize the number of genes in a line plot (Supplementary Text). Additionally, users can input a custom interval for each SNP (*e.g.,* based on linkage disequilibrium) by disabling the sliding window expansion.

For the second step of the algorithm, *cageminer* relies on the *module_enrichment()* function from the *BioNERO* package (Almeida-Silva and Venancio, 2021) to perform an enrichment analysis and find candidates from step 1 that co-occur in modules enriched in guide genes. Guides are genes known to be associated with the phenotype of interest, which can be passed as a single gene set in a character vector or as a 2-column data frame with gene IDs in the first column and gene classification (*e.g.,* Gene Ontology Terms or KEGG pathways) in the second column. In the latter case, *cageminer* will look for modules enriched in each class of guide genes rather than guides in general.

In the third step, the gene expression matrix used to infer the GCN is correlated to a binary matrix m_{ij} containing 1 if the sample *m* corresponds to the condition *j*, and 0 otherwise. This calculation, also known as gene significance, returns a point biserial correlation coefficient (r_{pb}) (Langfelder and Horvath, 2008) that indicates if genes have significantly increased or decreased expression levels in a particular condition. Further, as genes can be negative regulators of the phenotype of interest, negative correlation coefficients are also treated as biologically meaningful. Thus, the absolute value of r_{pb} is considered to define a gene significance threshold, as well as Student asymptotic *P*-values for correlation significance (by default, r_{pb} ≥0.2 and *P* <0.05).

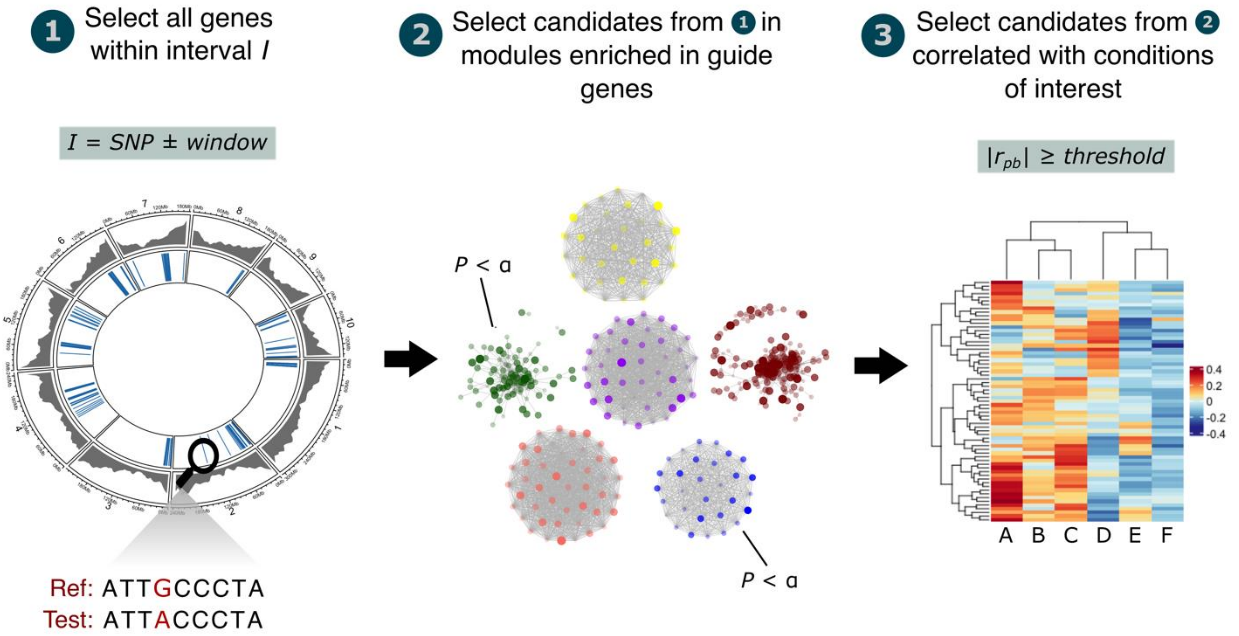

1 Select all genes within interval I
2 Select candidates from 1 in modules enriched in guide genes
3 Select candidates from 2 correlated with conditions of interest
I = SNP ± window
P < α
P < α
|r_pb| ≥ threshold
0.4
0.2
0
-0.2
-0.4
A B C D E F
Ref: ATTGCCCTA
Test: ATTACCCTA

Fig. 1. Summary of the *cageminer* algorithm. Candidate gene prioritization is performed in three sequential steps that can be run as a pipeline (recommended) or independently. The steps can be interpreted as different sources of evidence that candidates are causative genes. Thus, candidates that pass all three steps are considered high-confidence candidates.

2.2 Gene scoring

To score the prioritized candidate genes and further select the top *n* genes for validation, genes can be scored with the formula below:

$$S_i = r_{pb} \ \kappa$$

where

$\kappa = 2$ if the gene is a transcription factor

$\kappa = 2$ if the gene is a hub

$\kappa = 3$ if the gene is a hub and a transcription factor

$\kappa = 1$ if the gene is neither a hub nor a transcription factor

3 Application to a real dataset

A use case using RNA-seq on pepper (*Capsicum annuum*) response to Phytophthora root rot (Kim et al., 2018), as well as GWAS SNPs associated with resistance to Phytophthora root rot (Siddique et al., 2019) is available in the Supplementary Text. Pepper genes encoding transcription factors were downloaded from PlantTFDB 4.0 (Jin et al., 2017), and plant defense-related genes (MapMan annotations) were obtained from PLAZA Dicots 3.0 (Proost et al., 2015). From a list of 1265 putative candidates, *cageminer* identified 5 high-confidence candidate resistance genes (99.6% reduction). All candidates encode proteins related to known plant immunity-related processes (*e.g.*, immune signaling, oxidative stress, and lignan biosynthesis), supporting the effectiveness of the algorithm in finding biologically meaningful genes.

4 Conclusions

cageminer is the first R package to integrate GWAS-derived SNPs and gene coexpression networks to prioritize candidate genes involved in phenotypes of interest. This package will likely contribute to the advancement of population genomics and to the identification of genes for biotechnological applications.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.

Conflicts of interest: none declared.

REFERENCES

Almeida-Silva, F. and Venancio, T.M. (2021). BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction. bioRxiv: 2021.04.10.439287.

Baxter, I. (2020). We aren't good at picking candidate genes, and it's slowing us down. Curr. Opin. Plant Biol. **54**: 57–60.

Boudhrioua, C., Bastien, M., Torkamaneh, D., and Belzile, F. (2020). Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. BMC Plant Biol. **20**: 195.

Brodie, A., Azaria, J.R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Res. **44**: 6046–6054.

Buzanskas, M.E. et al. (2014). Genome-Wide Association for Growth Traits in Canchim Beef Cattle. PLoS One **9**: e94802.

Guo, J., Li, C., Zhang, X., Li, Y., Zhang, D., Shi, Y., Song, Y., Li, Y., Yang, D., and Wang, T. (2020). Transcriptome and GWAS analyses reveal candidate gene for seminal root length of maize seedlings under drought stress. Plant Sci. **292**.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res **45**: D1040–D1045.

Kim, M.S., Kim, S., Jeon, J., Kim, K.T., Lee, H.A., Lee, H.Y., Park, J., Seo, E., Kim, S.B., Yeom, S.I., Lee, Y.H., and Choi, D. (2018). Global gene expression profiling for fruit organs and pathogen infections in the pepper, *Capsicum annuum* L. Sci. Data **5**: 1–6.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**: 559.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating

Genomic Ranges. PLoS Comput. Biol. **9**: 1–10.

Maldonado Dos Santos, J.V., Ferreira, E.G.C., Passianotto, A.L.D.L., Brumer, B.B., Santos, A.B. Dos, Soares, R.M., Torkamaneh, D., Arias, C.A.A., Belzile, F., Abdelnoor, R.V., and Marcelino-Guimarães, F.C. (2019). Association mapping of a locus that confers southern stem canker resistance in soybean and SNP marker development. BMC Genomics **20**: 1–13.

Michno, J.M., Liu, J., Jeffers, J.R., Stupar, R.M., and Myers, C.L. (2020). Identification of nodulation-related genes in Medicago truncatula using genome-wide association studies and co-expression networks. Plant Direct **4**: 1–10.

Peat, G., Jones, W., Nuhn, M., Marugán, J.C., Newell, W., Dunham, I., and Zerbino, D. (2020). The open targets post-GWAS analysis pipeline. Bioinformatics: 1–2.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. **43**: D974–D981.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell **30**: 2922–2942.

Siddique, M.I., Lee, H.Y., Ro, N.Y., Han, K., Venkatesh, J., Solomon, A.M., Patil, A.S., Changkwian, A., Kwon, J.K., and Kang, B.C. (2019). Identifying candidate genes for Phytophthora capsici resistance in pepper (Capsicum annuum) via genotyping-by-sequencing-based QTL mapping and genome-wide association study. Sci. Rep. **9**: 1–15.

Wu, Y. et al. (2020). Multi-trait analysis for genome-wide association study of five psychiatric disorders. Transl. Psychiatry **10**: 209.

Yao, M. et al. (2020). GWAS and co-expression network combination uncovers multigenes with close linkage effects on the oleic acid content accumulation in Brassica napus. BMC Genomics **21**: 1–12.

CHAPTER 3:

Integration of genome-wide association studies and gene coexpression networks unveils promising soybean resistance genes against five common fungal pathogens

Chapter 3: Integration of genome-wide association studies and gene coexpression networks unveils promising soybean resistance genes against five common fungal pathogens

Fabricio Almeida-Silva[1*] and Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Research article

Situation: Accepted

Journal: Scientific Reports

Impact factor: 4.379; CiteScore: 7.1; Highest percentile: 93%

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

ABSTRACT

Soybean is one of the most important legume crops worldwide. However, soybean yield is dramatically affected by fungal diseases, leading to economic losses of billions of dollars yearly. Here, we integrated publicly available genome-wide association studies and transcriptomic data to prioritize candidate genes associated with resistance to *Cadophora gregata*, *Fusarium graminearum*, *Fusarium virguliforme, Macrophomina phaseolina*, and *Phakopsora pachyrhizi*. We identified 188, 56, 11, 8, and 3 high-confidence candidates for resistance to *F. virguliforme, F. graminearum, C. gregata, M. phaseolina* and *P. pachyrhizi*, respectively. The prioritized candidate genes are highly conserved in the pangenome of cultivated soybeans and are heavily biased towards fungal species-specific defense response. The vast majority of the prioritized candidate resistance genes are related to plant immunity processes, such as recognition, signaling, oxidative stress, systemic acquired resistance, and physical defense. Based on the number of resistance alleles, we selected the five most resistant accessions against each fungal species in the soybean USDA germplasm. Interestingly, the most resistant accessions do not reach the maximum theoretical resistance potential. Hence, they can be further improved to increase resistance in breeding programs or through genetic engineering. Finally, the coexpression network generated here is available in a user-friendly web application (https://soyfungigcn.venanciogroup.uenf.br/) and an R/Shiny package (https://github.com/almeidasilvaf/SoyFungiGCN) that serve as a public resource to explore soybean-pathogenic fungi interactions at the transcriptional level.

Keywords: plant immunity, QTL, systems biology, population genomics.

1 Introduction

Soybean (*Glycine max* (L.) Merr.) is a major legume crop worldwide, contributing to global food security and economy. However, soybean yield is significantly affected by diseases, with an estimated economic loss of 95.8 billion dollars from 1996 to 2006 in the US (Bandara et al., 2020). Most of the yield loss has been linked to foliar and stem/root diseases, which are mostly caused by phytopathogenic fungi (Bandara et al., 2020). Fungal diseases, such as sudden death syndrome, Fusarium wilt, brown stem rot and asian rust, can impact soybean crops through leaf damage, necrosis, chlorosis, and death (Pandey et al., 2011; Rincker et al., 2016; Bandara et al., 2020).

Over the past decade, several genome-wide association studies (GWAS) have uncovered multiple single-nucleotide polymorphisms (SNPs) associated with resistance to pathogenic fungi in soybean populations (Iquira et al., 2015; Sun et al., 2020; Kandel et al., 2018; Zhang et al., 2015; Rincker et al., 2016; Zhang et al., 2019; Chang et al., 2016). Nevertheless, GWAS often fail to accurately pinpoint the causative genes (Baxter, 2020). GWAS limitations are particularly challenging for self-pollinating plants (*e.g.,* soybean) because of limited recombination and strong linkage disequilibrium between causative and non-causative variants (Michno et al., 2020). Such limitations ultimately lead to large genetic intervals with several genes, hindering causative gene identification. Because of the exponential accumulation of genomic and transcriptomic data in public databases (Schwartz, 2020; Deshmukh et al., 2014; Schaefer et al., 2018; Baker et al., 2019; Wen et al., 2018), integrative analyses to prioritize candidate genes have become a promising approach. This strategy consists in investigating the transcriptional patterns of all the genes near a significant SNP. Hence, the combination of multiple sources of evidence can result in richer and narrower sets of high-confidence candidate genes for downstream experimental validation towards biotechnological applications.

Here, we integrated multiple publicly available RNA-seq and GWAS datasets to identify high-confidence candidate genes for resistance to five phytopathogenic fungi. The prioritized resistance genes are species-specific and highly conserved in the pangenome of cultivated soybeans. The candidate resistance genes against each species are involved in various immunity-related processes, such as recognition, signaling, oxidative stress, and apoptosis. Finally, we highlighted the five most resistant accessions against each fungal species in the USDA germplasm, uncovering important information for breeding programs and genetic engineering initiatives. Finally, the coexpression network resulting from this work was also made available as a publicly available web application (https://soyfungigcn.venanciogroup.uenf.br/) and R/Shiny package (https://github.com/almeidasilvaf/SoyFungiGCN).

2 Materials and Methods

2.1 Curation of resistance-associated SNPs

SNPs that contribute to resistance against phytopathogenic fungi were manually curated from the scientific literature (Table 1; Supplementary Table S1). SNPs that were identified using the Gmax_a1.v1 genome were converted to their corresponding sites in the Gmax_a2.v1 assembly using the .vcf files for both assemblies available at Soybase (Brown et al., 2020).

Table 1. GWAS included in this work.

Reference	Pathogen	Resistance SNPs
(Zhang et al., 2019)	*F. graminearum*	12
(Bao et al., 2015)	*F. virguliforme*	8
(Chang et al., 2016)	*C. gregata / F. virguliforme / P. pachyrhizi*	2 /1 /2
(Zhang et al., 2015)	*F. virguliforme*	32
(Swaminathan et al., 2019)	*F. virguliforme*	27
(Vinholes et al., 2019)	*M. phaseolina*	4
(Coser et al., 2017)	*M. phaseolina*	12
(Rincker et al., 2016)	*C. gregata*	7

2.2 Transcriptome data

Gene expression estimates in transcripts per million mapped reads (TPM, Kallisto estimation) were retrieved from the Soybean Expression Atlas (Machado et al., 2020). Additional RNA-seq samples comprising soybean tissues infected with fungal pathogens were retrieved from a recent publication from our group (Almeida-Silva and Venancio, 2021c). We filtered the SNP and transcriptome datasets to keep only fungal species that were represented by both data sources. A total of 150 RNA-seq samples from soybean tissues infected with fungal pathogens were selected (Supplementary Table S2). Finally, genes with median expression values lower than 5 were excluded to attenuate noise, resulting in an 18748 *x* 150 gene expression matrix for downstream analyses.

2.3 Selection of guide genes

MapMan annotations for soybean genes were retrieved from the PLAZA 3.0 Dicots database (Proost et al., 2015). Genes assigned to defense-related pathways (*e.g.,* pathogenesis-related proteins, lignin biosynthesis, oxidative stress, and phytohormone regulation) were used as guides (Supplementary Table S3).

2.4 Candidate gene mining and functional analyses

Gene expression data were adjusted for confounding artifacts and quantile normalized with the R package BioNERO (Almeida-Silva and Venancio, 2021a). An unsigned coexpression network was inferred with BioNERO using Pearson's r as correlation. All genes located in a 2 Mb sliding window relative to each SNP were selected as putative candidates, as previously proposed (Brodie et al., 2016). Candidate genes were prioritized using the algorithm implemented in the R package cageminer (Almeida-Silva and Venancio, 2021b), with an r_{pb} threshold of 0.2 for gene significance (gene-trait correlation). Enrichment analyses were also performed with BioNERO, using functional annotations from the PLAZA 4.0 database (Van Bel et al., 2018). To rank the prioritized candidates, they were given scores using the formula:

$$S = r_{pb}\kappa$$

where
r_{pb}= point-biserial correlation coefficient (cageminer algorithm)
$\kappa = 2$ if the gene is a transcription factor
$\kappa = 2$ if the gene is a hub
$\kappa = 3$ if the gene is a hub and a transcription factor
$\kappa = 1$ if the gene is neither a hub nor a transcription factor

2.5 Selection of most resistant accessions from the USDA germplasm

The VCF file with genotypic information for all accessions in the USDA germplasm was downloaded from Soybase (Brown et al., 2020). Scores 0, 1, and 2 were attributed to accessions with 0, 1, and 2 beneficial SNPs (effect size >0), respectively, whereas scores 2, 1, and 0 were attributed to accessions with 0, 1, and 2 deleterious SNPs (effect size <0). The resistance potential of the best accessions was calculated as a ratio of the attributed scores to the theoretical maximum score (all beneficial SNPs and no deleterious SNPs).

3 Results and discussion

3.1 Data summary and genomic distribution of SNPs

After filtering the datasets to keep only fungal species represented by both SNP and transcriptome information, we kept five common phytopathogenic fungi: *Cadophora gregata*, *Fusarium graminearum*, *Fusarium virguliforme*, *Macrophomina phaseolina*, and *Phakopsora pachyrhizi* (Figure 1A). Overall, SNPs were located in gene-rich regions of the genome (Figure 1B). SNPs were unevenly distributed across chromosomes, except for *F. virguliforme* (Figure 1C). Further, we found that most SNPs were located in intergenic regions (Figure 1D). Hence, predicting SNP effect on genes would not be suitable for this trait.

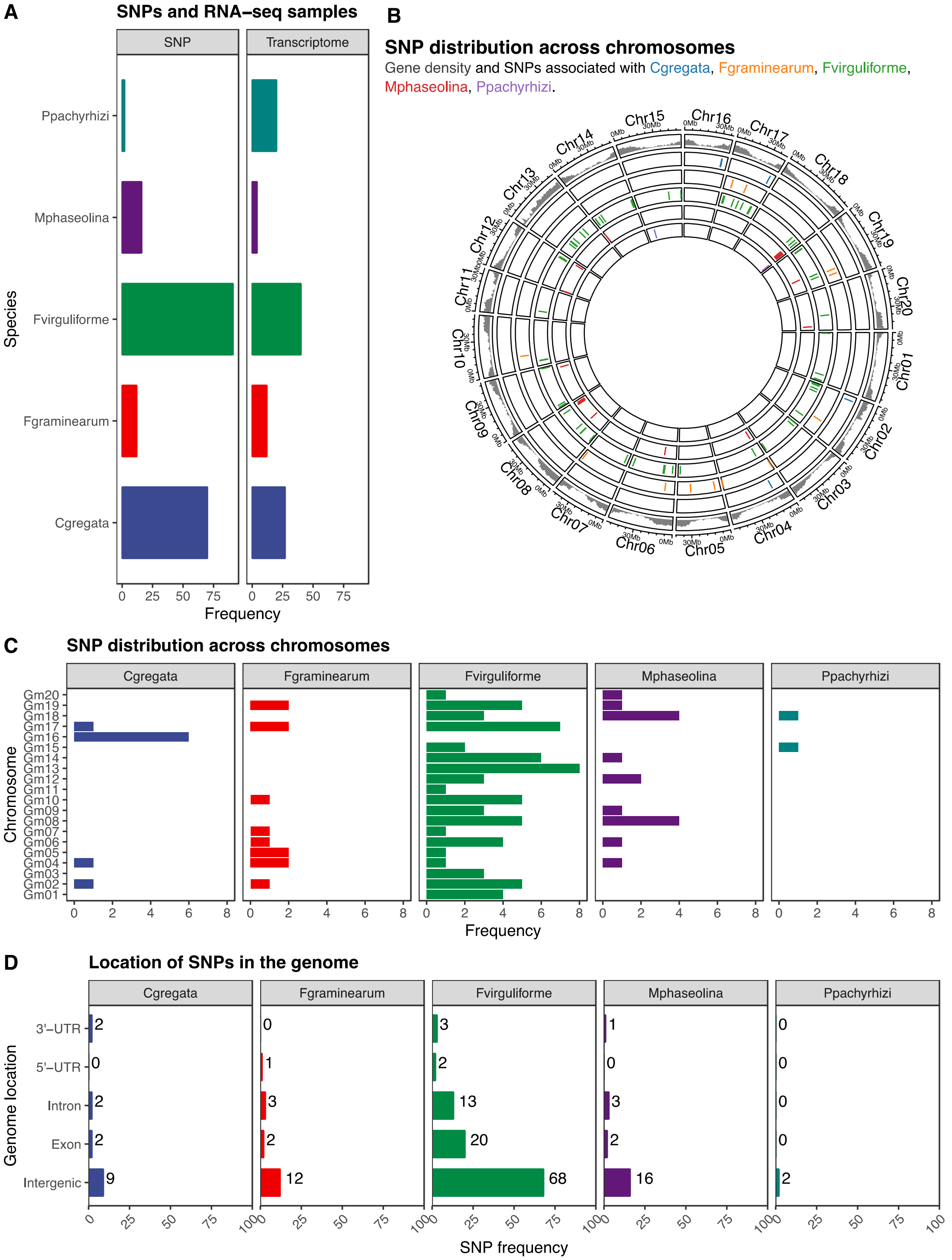


Figure 1. Data summary and genomic distribution of SNPs. A. Frequency of SNPs and RNA-seq samples included in this study. B. Genomic coordinates of resistance SNPs against each fungal pathogen. The outer track represents gene density, whereas inner tracks represent the SNP positions for each species. C. SNP distribution across chromosomes. Overall, there is an uneven distribution of SNPs across chromosomes. D. Genomic location of SNPs. Most SNPs are located in intergenic regions.

3.2 Candidate gene mining reveals a highly species-specific immune response

Using defense-related genes as guides, the cageminer algorithm identified 188, 56, 11, 8, and 3 high-confidence genes for *F. virguliforme, F. graminearum, C. gregata, M. phaseolina*, *and P. pachyrhizi*, respectively (Figure 2)*.* Only three genes were shared between species, revealing a high specificity in plant-pathogen interactions for these species. The three genes are shared by *F. virguliforme* and *F. graminearum*, suggesting that some conservation can occur at the genus level, but not at other broader taxonomic levels.

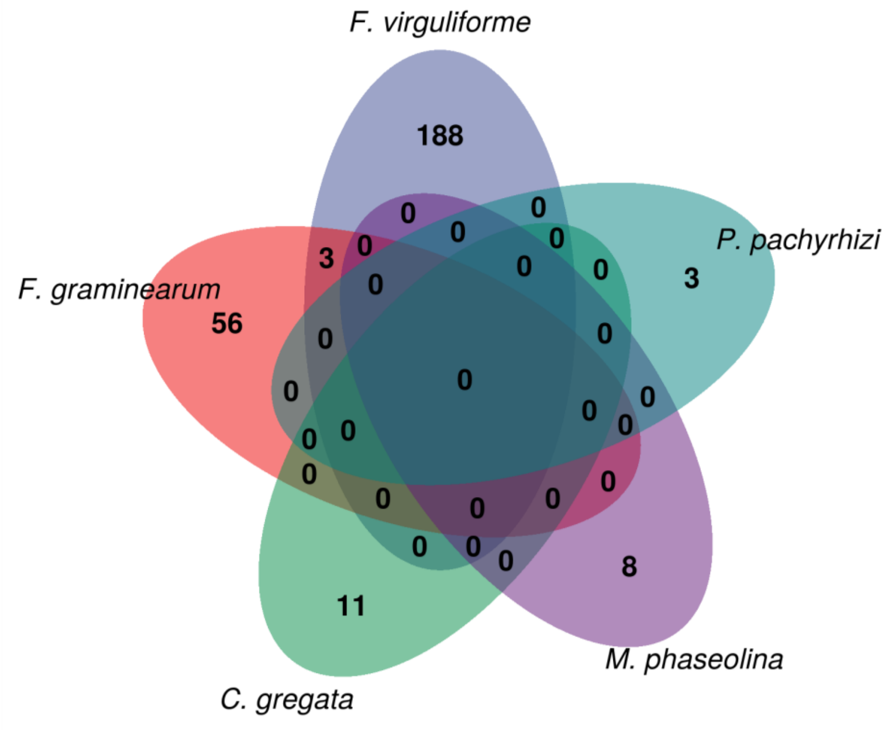


Figure 2. Venn diagram of prioritized candidate resistance genes against each species. The diagram demonstrates a high species-specific response to each pathogen, as genes are mostly not shared. Only three genes are shared between *F. graminearum* and *F. virguliforme*, suggesting some conservation at the genus level.

The specificity of resistance genes to particular species has been widely reported (Kourelis and Van Der Hoorn, 2018; Ning and Wang, 2018; Li et al., 2020; Durrant and Dong, 2004). This phenomenon imposes a challenge for biotechnological applications, as it requires pyramiding many different genes to render elite cultivars resistant to different pathogens. However, we cannot rule out that the species-specific trend we observed results from low diversity in the association panels in the GWAS we analyzed. Additionally, as SNP and transcriptome data are not available for multiple pathogen strains, we might overlook broad-spectrum resistance genes that confer resistance to multiple strains of the same species (Ning and Wang, 2018).

Further, we manually curated the high-confidence candidate resistance genes to predict the putative role of their products in plant immunity (Supplementary Table S4). Most of the prioritized candidates (28%) encode proteins involved in immune signaling, although it does not apply to all fungi species (Figure 3). Candidates also encode proteins that play a role in recognition, phytohormone metabolism, systemic acquired resistance, transport, transcriptional regulation, oxidative stress, apoptosis, physical defense, and direct function against fungi (Figure 3). Interestingly, 21 candidate genes lack functional description and, hence, we could not infer their roles in plant immunity (*n*=2, 4, 14, and 1 for *C. gregata, F. virguliforme,* and *P. pachyrhizi*, respectively). Nevertheless, as they were identified as high-confidence candidate genes, we hypothesize that they encode defense-related proteins. We also developed a scheme that was used to rank high-confidence candidate genes, which can be used to prioritize candidates for experimental validation in future studies (Table 2).

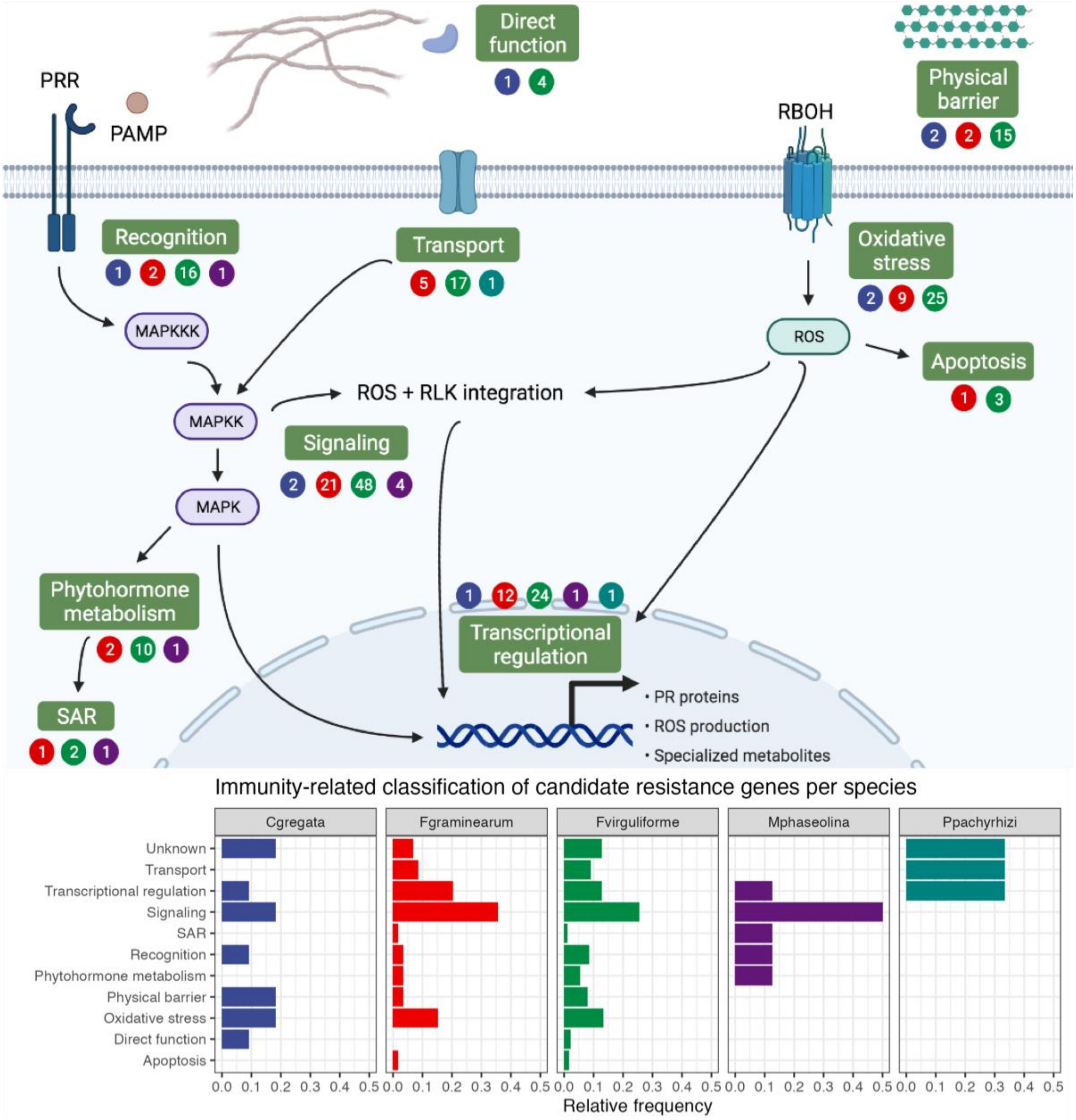


Figure 3. Prioritized candidate resistance genes and their putative role in plant immunity. Numbers in circles represent absolute frequencies of resistance genes against *C. gregata* (blue)*, F. graminearum* (red)*, F. virguliforme* (green)*, M. phaseolina* (purple)*,* and *P pachyrhizi* (turquoise). PRR, pattern recognition receptor. PAMP, pathogen-associated molecular pattern. MAPKKK, mitogen-activated protein kinase kinase kinase. MAPKK, mitogen-activated protein kinase kinase. MAPK, mitogen-activated protein kinase. SAR, systemic acquired resistance. RBOH, respiratory burst oxidase homolog. ROS, reactive oxygen species. RLK, receptor-like kinase. PR, pathogenesis-related. Figure designed with Biorender (biorender.com).

Table 2. Top 10 candidate resistance genes against each fungal species and their putative roles in plant immunity. The predicted function for each gene was manually curated from the description of the best ortholog in *Arabidopsis thaliana*, using functional annotations from Soybase and TAIR.

Gene	Predicted function	Resistance to	Role
Glyma.16G170100	Cell wall biogenesis-related extensin 3	*C. gregata*	Physical barrier
Glyma.02G026700	Transcriptional repressor SIN3	*C. gregata*	Transcriptional regulation
Glyma.02G026900	Galacturonosyltransferase	*C. gregata*	Physical barrier
Glyma.02G029300	SAM domain-containing	*C. gregata*	Unknown
Glyma.16G155100	Aquaporin	*C. gregata*	Oxidative stress
Glyma.17G217000	Class V chitinase	*C. gregata*	Direct function
Glyma.17G213600	Calcium-binding EF hand	*C. gregata*	Signaling
Glyma.17G231800	Clathrin adaptor EPSIN1	*C. gregata*	Recognition
Glyma.02G047000	Thiosulfate sulfurtransferase/rhodanese	*C. gregata*	Oxidative stress
Glyma.16G150500	Unknown	*C. gregata*	Unknown
Glyma.17G087500	SOUL heme-binding protein	*F. graminearum*	Oxidative stress
Glyma.06G121300	GRAS transcription factor	*F. graminearum*	Transcriptional regulation
Glyma.05G070300	Tobamovirus multiplication 2A	*F. graminearum*	Recognition
Glyma.04G013500	BURP domain-containing protein	*F. graminearum*	Physical barrier
Glyma.06G105000	ERF/AP2 transcription factor	*F. graminearum*	Transcriptional regulation
Glyma.05G062400	2OG-Fe(II) oxygenase	*F. graminearum*	Oxidative stress
Glyma.05G063600	ERF/AP2 transcription factor	*F. graminearum*	Transcriptional regulation
Glyma.05G115700	RING domain ubiquitin E3 ligase	*F. graminearum*	Signaling
Glyma.17G116100	MAPK signaling-related protein	*F. graminearum*	Signaling
Glyma.05G103600	Peroxidase	*F. graminearum*	Oxidative stress
Glyma.13G081000	Nodulin-like amino acid transporter	*F. virguliforme*	Transport
Glyma.01G225600	Unknown	*F. virguliforme*	Unknown
Glyma.02G210500	bHLH transcription factor	*F. virguliforme*	Transcriptional regulation
Glyma.01G162500	BIG1 protein	*F. virguliforme*	Apoptosis
Glyma.17G061400	Peroxidase	*F. virguliforme*	Oxidative stress
Glyma.19G010100	HD-Zip transcription factor	*F. virguliforme*	Transcriptional regulation
Glyma.18G276800	Amino acid transporter	*F. virguliforme*	Oxidative stress
Glyma.05G209900	PLAC8 family protein	*F. virguliforme*	Apoptosis
Glyma.14G025100	Inositol-1,4,5-trisphosphate 5-phosphatase	*F. virguliforme*	Signaling
Glyma.19G117800	Unknown	*F. virguliforme*	Unknown
Glyma.20G203900	Type I serine/threonine protein phosphatase	*M. phaseolina*	Signaling

Glyma.08G316500	Calmodulin-dependent protein kinase	*M. phaseolina*	Signaling
Glyma.06G187200	R-gene-mediated resistance, lipase	*M. phaseolina*	SAR
Glyma.09G218600	Cytochrome P450, family 707, subfamily A	*M. phaseolina*	Phytohormone metabolism
Glyma.09G216800	Pectin acetylesterase	*M. phaseolina*	Signaling
Glyma.20G216600	Dof-type transcription factor	*M. phaseolina*	Transcriptional regulation
Glyma.08G332800	Calcineurin B-like calcium sensor	*M. phaseolina*	Signaling
Glyma.18G301700	Leucine-rich repeat receptor kinase (LRR-RK)	*M. phaseolina*	Recognition
Glyma.15G125900	Magnesium transporter CorA-like	*P. pachyrhizi*	Transport
Glyma.18G286900	Unknown	*P. pachyrhizi*	Unknown
Glyma.15G123900	CBF1 interacting co-repressor CIR	*P. pachyrhizi*	Transcriptional regulation

3.3 Pangenome presence/absence variation analysis demonstrates that most prioritized genes are core genes

We analyzed PAV patterns for our prioritized candidate genes in the recently published pangenome of cultivated soybeans to unveil which soybean genotypes contain prioritized candidate genes and explore gene presence/absence variation patterns across genomes (Torkamaneh et al., 2021). We found that most candidates are present in all 204 accessions (Supplementary Figure 1A). This trend is not surprising, as the gene content in this pangenome is highly conserved, with ~91% of the genes being shared by >99% of the genomes. Although the variable genome is enriched in genes associated with defense, signaling, and plant development, this trend was not found in our gene set.

Further, we investigated if gene PAV patterns could be explained by the geographical origins of the accessions (Supplementary Figure 1B). Strikingly, we observed no clustering by geographical origin, suggesting that gene PAV is not affected by population structure. As this pangenome is comprised of improved soybean accessions (Torkamaneh et al., 2021), the lack of population structure effect can be due to breeding programs targeting optimal adaptation to different environmental conditions (*e.g.,* latitude and climate), even if they are in the same country.

3.4 Screening of the USDA germplasm reveals a room for genetic improvement

We inspected the USDA germplasm to find the top 5 most resistant genotypes against each fungal pathogen (see Materials and Methods for details). Strikingly, the most resistant genotypes do not contain all resistance alleles, revealing that, theoretically, they could be further improved to increase resistance (Table 3). All resistance-associated SNPs against *P. pachyrhizi* are present in some accessions, but this is because only two SNPs have been reported for this species. Additionally, none of the reported SNPs for *F. graminearum* have been identified in the SoySNP50k collection. Hence, we could not predict the most resistant accessions to this fungal species in the USDA germplasm.

Although some individual genes can confer full race-specific resistance to some pathogens, their durability in the field is often short because of pathogen evolution (Ning and Wang, 2018). Thus, pyramiding quantitative trait loci (QTL) that confer partial resistance has been proposed a strategy to confer long-term resistance (Li et al., 2020).

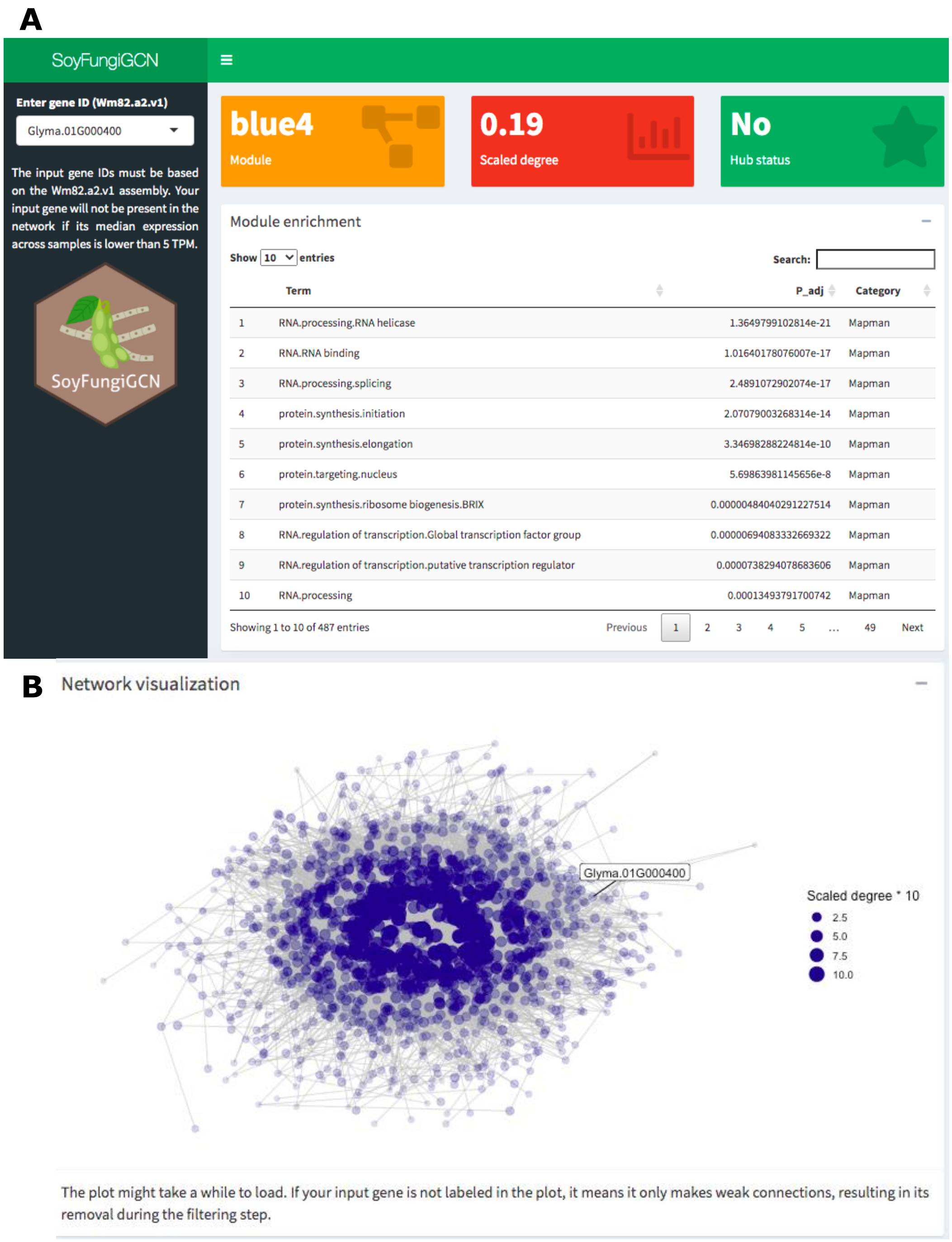
To accomplish this, the most resistant genotypes identified here can be targets of allele pyramiding in breeding programs using marker-assisted selection. Alternatively, these genotypes might have their genomes edited with CRISPR/Cas systems to introduce beneficial alleles or remove deleterious alleles, ultimately boosting resistance.

Table 3. Top 5 most resistant soybean accessions against each fungal pathogen. Overall, the best genotypes do not reach the maximum potential. An exception is observed for *P. pachyrhizi*-resistant genotypes, but this is likely due to the small number of resistance SNPs. None of the resistance SNPs for *F. graminearum* have been identified in the USDA SoySNP50k compendium and, hence, we could not predict resistance potential against this species.

Accession	Score	Potential	Species
PI594466	102	0.73	*C. gregata*
PI578477A	100	0.71	*C. gregata*
PI437571	100	0.71	*C. gregata*
PI567520A	100	0.71	*C. gregata*
PI274507	100	0.71	*C. gregata*
PI339871C	82	0.60	*F. virguliforme*
PI378694	80	0.59	*F. virguliforme*
PI407145	80	0.59	*F. virguliforme*
PI424107A	80	0.59	*F. virguliforme*
PI479753A	80	0.59	*F. virguliforme*
PI594760B	24	0.75	*M. phaseolina*
PI479752	24	0.75	*M. phaseolina*
PI603706A	24	0.75	*M. phaseolina*
PI603531A	24	0.75	*M. phaseolina*
PI603412A	24	0.75	*M. phaseolina*
PI603547	4	1	*P. pachyrhizi*
PI639559A	4	1	*P. pachyrhizi*
PI639559B	4	1	*P. pachyrhizi*
PI326582A	4	1	*P. pachyrhizi*
PI407057	4	1	*P. pachyrhizi*

3.5 Development of a user-friendly web application for network exploration

To facilitate network exploration and data reuse, we developed a user-friendly web application named SoyFungiGCN (https://soyfungigcn.venanciogroup.uenf.br/). Users can input a soybean gene of interest (Wm82.a2.v1 assembly) and visualize the gene's module, scaled intramodular degree, and hub status (Figure 4A). Additionally, users can explore enriched GO terms, Mapman bins and/or Interpro domains associated with the input gene's module (Figure 4A). Users can also visualize a network plot with the input gene and its coexpression neighbors (Figure 4B). This resource can be particularly useful for researchers studying soybean response to other fungal species, as they can check if their genes of interest are located in defense-related coexpression modules. Also, researchers studying other species can verify if the soybean ortholog of their genes of interest is located in a defense-related module. The application is also available as an R package named SoyFungiGCN (https://github.com/almeidasilvaf/SoyFungiGCN). This package lets users run the application locally as a Shiny app, ensuring the application will always be available, even in case of server downtime.



Enter gene ID (Wm82.a2.v1)

Glyma.01G000400

The input gene IDs must be based on the Wm82.a2.v1 assembly. Your input gene will not be present in the network if its median expression across samples is lower than 5 TPM.

Module enrichment

Show 10 entries

Search:

	Term	P_adj	Category
1	RNA.processing.RNA helicase	1.3649799102814e-21	Mapman
2	RNA.RNA binding	1.01640178076007e-17	Mapman
3	RNA.processing.splicing	2.4891072902074e-17	Mapman
4	protein.synthesis.initiation	2.07079003268314e-14	Mapman
5	protein.synthesis.elongation	3.34698288224814e-10	Mapman
6	protein.targeting.nucleus	5.69863981145656e-8	Mapman
7	protein.synthesis.ribosome biogenesis.BRIX	0.00000484040291227514	Mapman
8	RNA.regulation of transcription.Global transcription factor group	0.00000694083332669322	Mapman
9	RNA.regulation of transcription.putative transcription regulator	0.0000738294078683606	Mapman
10	RNA.processing	0.00013493791700742	Mapman

Showing 1 to 10 of 487 entries

Previous 1 2 3 4 5 … 49 Next

The plot might take a while to load. If your input gene is not labeled in the plot, it means it only makes weak connections, resulting in its removal during the filtering step.

Figure 4. Functionalities in the SoyFungiGCN web application. A. Screenshot of the page users see when they access the application. In the sidebar, users can specify the ID of a gene of interest (Wm82.a2.v1 assembly). For each gene, users can see the gene's module (orange box), scaled degree (red box), hub gene status (green box), and an interactive table with enrichment results for MapMan bins, Interpro domains and Gene Ontology terms associated the gene's module. P-values from enrichment results are adjusted for multiple testing with Benjamini-Hochberg correction. B. Network visualization plot. Users can optionally visualize the input gene and its position in the module by clicking the plus (+) icon in the "Network visualization" tab below the enrichment table. As the plot can take a few seconds to render (~2-5 seconds), it is hidden by default.

4 Conclusions

By integrating publicly available GWAS and RNA-seq data, we found promising candidate genes in soybean associated with resistance to five common phytopathogenic fungi, namely *C. gregata*, *F. graminearum*, *F. virguliforme*, *M. phaseolina,* and *P. pachyrhizi.* The prioritized candidates encode proteins that play a role immunity-related processes such as in recognition, signaling, transcriptional regulation, oxidative stress, and physical defense. We have also found the top 5 most resistant soybean accessions against each fungal species and hypothesize that they can be further genetically improved in breeding programs with marker-assisted selection or through genome editing. The coexpression network generated here was also made available in a web resource and R package to help in future studies on soybean-pathogenic fungi interactions.

Data availability

All data and code used in this study are available in our GitHub repository (https://github.com/almeidasilvaf/SoyFungi_GWAS_GCN) to ensure full reproducibility.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.

Author contributions

Conceived the study: FA-S and TMV. Data analysis: FA-S. Funding, project coordination and infrastructure: TMV. Manuscript writing: FA-S and TMV.

REFERENCES

Almeida-Silva, F. and Venancio, T.M. (2021a). BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction. bioRxiv: 2021.04.10.439287.

Almeida-Silva, F. and Venancio, T.M. (2021b). cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks. bioRxiv: 1–9.

Almeida-Silva, F. and Venancio, T.M. (2021c). Pathogenesis-related protein 1 (PR-1) genes in soybean: genome-wide identification, structural analysis and expression profiling under multiple biotic and abiotic stresses. bioRxiv **1**: 1–23.

Baker, R.L., Leong, W.F., Brock, M.T., Rubin, M.J., Markelz, R.J.C., Welch, S., Maloof, J.N., and Weinig, C. (2019). Integrating transcriptomic network reconstruction and eQTL analyses reveals mechanistic connections between genomic architecture and Brassica rapa development. PLOS Genet. **15**: e1008367.

Bandara, A.Y., Weerasooriya, D.K., Bradley, C.A., Allen, T.W., and Esker, P.D. (2020). Dissecting the economic impact of soybean diseases in the United States over two decades. PLoS One **15**: 1–28.

Bao, Y., Kurle, J.E., Anderson, G., and Young, N.D. (2015). Association mapping and genomic prediction for resistance to sudden death syndrome in early maturing soybean germplasm. Mol. Breed. **35**: 1–14.

Baxter, I. (2020). We aren’t good at picking candidate genes, and it’s slowing us down. Curr. Opin. Plant Biol. **54**: 57–60.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van De Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46**: D1190–D1196.

Brodie, A., Azaria, J.R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Res. **44**: 6046–6054.

Brown, A. V, Conners, S.I., Huang, W., Wilkey, A.P., Grant, D., Weeks, N.T., Cannon, S.B., Graham, M.A., and Nelson, R.T. (2020). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. **13**: 1–6.

Chang, H.X., Lipka, A.E., Domier, L.L., and Hartman, G.L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Phytopathology **106**: 1139–1151.

Coser, S.M., Reddy, R.V.C., Zhang, J., Mueller, D.S., Mengistu, A., Wise, K.A., Allen, T.W., Singh, A., and Singh, A.K. (2017). Genetic architecture of charcoal rot (Macrophomina phaseolina) resistance in soybean revealed using a diverse panel. Front. Plant Sci. **8**: 1–12.

Deshmukh, R., Sonah, H., Patil, G., Chen, W., Prince, S., Mutava, R., Vuong, T., Valliyodan, B., and Nguyen, H.T. (2014). Integrating omic approaches for abiotic stress tolerance in soybean. Front. Plant Sci. **5**: 1–12.

Durrant, W.E. and Dong, X. (2004). Systemic acquired resistance. Annu. Rev. Phytopathol. **42**: 185–209.

Iquira, E., Humira, S., and François, B. (2015). 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 Biol. **15**: 1–12.

Kandel, R., Chen, C.Y., Grau, C.R., Dorrance, A.E., Liu, J.Q., Wang, Y., and Wang, D. (2018). Soybean resistance to white mold: Evaluation of soybean germplasm under different conditions and validation of QTL. Front. Plant Sci. **9**: 1–12.

Kourelis, J. and Van Der Hoorn, R.A.L. (2018). Defended to the Nines: 25 years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. Plant Cell.

Li, W., Deng, Y., Ning, Y., He, Z., and Wang, G.L. (2020). Exploiting Broad-Spectrum Disease Resistance in Crops: From Molecular Dissection to Breeding. Annu. Rev. Plant Biol. **71**: 575–603.

Machado, F.B., Moharana, K.C., Almeida-Silva, F., Gazara, R.K., Pedrosa-Silva, F., Coelho, F.S., Grativol, C., and Venancio, T.M. (2020). Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas. Plant J.: 0–2.

Michno, J.M., Liu, J., Jeffers, J.R., Stupar, R.M., and Myers, C.L. (2020). Identification of nodulation-related genes in Medicago truncatula using genome-wide association studies and co-expression networks. Plant Direct **4**: 1–10.

Ning, Y. and Wang, G.L. (2018). Breeding plant broad-spectrum resistance without yield penalties. Proc. Natl. Acad. Sci. U. S. A. **115**: 2859–2861.

Pandey, A.K., Yang, C., Zhang, C., Graham, M.A., Horstman, H.D., Lee, Y., Zabotina, O.A., Hill, J.H., Pedley, K.F., and Whitham, S.A. (2011). Functional analysis of the asian soybean rust resistance pathway mediated by Rpp2. Mol. Plant-Microbe Interact. **24**: 194–206.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics.

Nucleic Acids Res. **43**: D974–D981.

Rincker, K., Lipka, A.E., and Diers, B.W. (2016). Genome-Wide Association Study of Brown Stem Rot Resistance in Soybean across Multiple Populations. Plant Genome **9**: plantgenome2015.08.0064.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell **30**: 2922–2942.

Schwartz, T.S. (2020). The promises and the challenges of integrating multi-omics and systems biology in comparative stress biology. Integr. Comp. Biol. **53**: 1689–1699.

Sun, M., Jing, Y., Zhao, X., Teng, W., Qiu, L., Zheng, H., Li, W., and Han, Y. (2020). Genome-wide association study of partial resistance to sclerotinia stem rot of cultivated soybean based on the detached leaf method. PLoS One **15**: 1–15.

Swaminathan, S., Das, A., Assefa, T., Knight, J.M., Da Silva, A.F., Carvalho, J.P.S., Hartman, G.L., Huang, X., Leandro, L.F., Cianzio, S.R., and Bhattacharyya, M.K. (2019). Genome wide association study identifies novel single nucleotide polymorphic loci and candidate genes involved in soybean sudden death syndrome resistance. PLoS One **14**: 1–21.

Torkamaneh, D., Lemay, M.-A., and Belzile, F. (2021). The Pan-genome of the Cultivated Soybean (PanSoy) Reveals an Extraordinarily Conserved Gene Content. Plant Biotechnol. J. **n/a**.

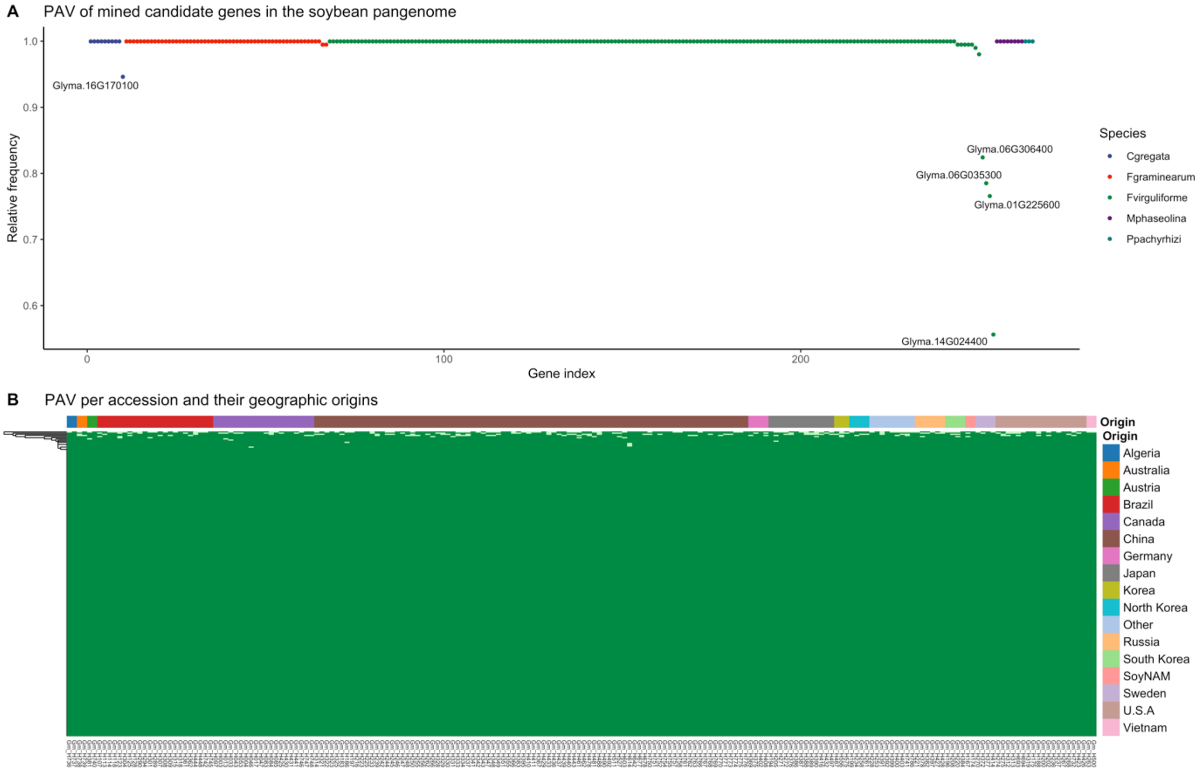
Vinholes, P., Rosado, R., Roberts, P., Borém, A., and Schuster, I. (2019). Single nucleotide polymorphism-based haplotypes associated with charcoal rot resistance in Brazilian soybean germplasm. Agron. J. **111**: 182–192.

Wen, Z. et al. (2018). Integrating GWAS and gene expression data for functional characterization of resistance to white mould in soya bean. Plant Biotechnol. J. **16**: 1825–1835.

Zhang, C., Zhao, X., Qu, Y., Teng, W., Qiu, L., Zheng, H., Wang, Z., Han, Y., and Li, W. (2019). Loci and candidate genes in soybean that confer resistance to Fusarium graminearum. Theor. Appl. Genet. **132**: 431–441.

Zhang, J., Singh, A., Mueller, D.S., and Singh, A.K. (2015). Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. Plant J. **84**: 1124–1136.

Supplementary Figures



Supplementary Figure 1. Presence/absence variation (PAV) of prioritized candidate genes in the soybean pangenome. A. Relative frequency of accessions containing each candidate gene. Most candidates are present in all accessions. Candidate genes with lower frequency in the pangenome are labeled. B. PAV per accessions and their geographic distribution. The patterns of gene PAV cannot be explained by the geographic origins of the accessions.

CHAPTER 4:

Discovering and prioritizing candidate resistance genes against soybean pests by integrating GWAS and gene coexpression networks

Chapter 4: Discovering and prioritizing candidate resistance genes against soybean pests by integrating GWAS and gene coexpression networks

Fabricio Almeida-Silva[1*] and Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Research article

Situation: In preparation

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com.

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com.

ABSTRACT

Soybean is one of the most important legume crops worldwide. However, soybean pests are responsible for severe economic losses because of reduced crop yield. Here, we integrated publicly available genome-wide association studies and transcriptomic data to prioritize candidate resistance genes against the insects *Aphis glycines* and *Spodoptera litura*, and the nematode *Heterodera glycines*. We identified 171, 7, and 228 high-confidence candidate resistance genes against *A. glycines, S. litura,* and *H. glycines*, respectively. We found some overlap of candidate genes between insect species, but not between insects and *H. glycines*. Although 15% of the prioritized candidate genes encode proteins of unknown function, the vast majority of the candidates are related to plant immunity processes, such as transcriptional regulation, signaling, oxidative stress, recognition, and physical defense. Based on the number of resistance alleles, we selected the ten most promising accessions against each pest species in the soybean USDA germplasm. The most resistant accessions do not reach the maximum theoretical resistance potential, indicating that they can be further improved to increase resistance in breeding programs or through genetic engineering. Finally, the coexpression networks generated here are available in a user-friendly web application (https://soypestgcn.venanciogroup.uenf.br/) and an R/Shiny package (https://github.com/almeidasilvaf/SoyPestGCN) that serve as a public resource to explore soybean-pest interactions at the transcriptional level.

Keywords: plant immunity, QTL, gene discovery, population genomics.

1 Introduction

Soybean (*Glycine max* (L.) Merr.) is the world's main legume crop, with a primary impact in human and animal nutrition, and in industrial applications. However, soybean fields are significantly affected by pests (insects and nematodes) that lead to dramatic yield losses. The major pests in soybean fields are the soybean aphid (*Aphis glycines* Matsumura) and the soybean cyst nematode (*Heterodera glycines* Ichinohe), which are responsible for annual losses of US$4 billion and US$4.5 billion in the US, respectively (Bandara et al., 2020; Koenning and Wrather, 2010). In Brazil, the world's leading soybean producer, insect pests reduce yield by 7.7%, which corresponds to an economic loss of US$ 17.7 billion (Oliveira et al., 2014).

Over the past few years, many genome-wide associations studies (GWAS) have been performed to identify single-nucleotide polymorphisms (SNPs) associated with soybean resistance to insect and nematode pests (Liu et al., 2019; Hanson et al., 2018; Zhao et al., 2017; Bao et al., 2014; Natukunda et al., 2019). However, as GWAS typically cannot accurately pinpoint causative genes, multi-omics data integration has helped predict high-confidence candidate genes associated with traits of interest (Baxter, 2020; Michno et al., 2020). Recently, we identified high-confidence candidate genes against fungal diseases using cageminer, a graph-based algorithm recently developed by our group to integrate GWAS and transcriptomic data to prioritize candidate genes (Almeida-Silva and Venancio, 2021b, 2021c). Thus, we hypothesize that our algorithm can also reveal high-confidence candidate genes that can be used to engineer soybean lines with increased resistance to pests.

Here, we integrated multiple publicly available RNA-seq and GWAS datasets to identify high-confidence candidate genes associated with resistance to pests. We found a high overlap of resistance genes between insects, but not between insects and nematodes, suggesting that these classes trigger different defense responses. The candidate resistance genes against each species are involved in several immunity-related processes, such as transcriptional regulation, signaling, oxidative stress, recognition, and phytohormone metabolism. Strikingly, 15% of the candidates encode proteins of unknown function, revealing a hidden catalog of potential resistance genes. Finally, we highlighted the ten most resistant accessions against each pest species in the USDA germplasm, uncovering important information for breeding programs and genetic engineering initiatives. The coexpression networks resulting from this work were also made available as a web application (https://soypestgcn.venanciogroup.uenf.br/) and R/Shiny package (https://github.com/almeidasilvaf/SoyPestGCN).

2 Materials and Methods

2.1 Curation of resistance-associated SNPs and pan-genome data

SNPs with significant association to resistance against soybean pests were manually curated from published GWAS data (Table 1; Supplementary Table S1). SNPs that were present in the SoySNP50k database were identified with their standard nomenclature, and the VCF file for the SoySNP50k was downloaded from Soybase (Brown et al., 2020). Additionally, a matrix of gene presence/absence variation (PAV) in the pan-genome of cultivated soybeans (*n* = 204 genomes from 24 countries and 5 continents) (Torkamaneh et al., 2021) was also used.

Table 1. GWAS included in this work. *N*, number of significant resistance-related SNPs in each study.

Reference	Organism	*N*
(Liu et al., 2019)	*H. glycines*	11
(Liu et al., 2016)	*S. litura*	6
(Zhang et al., 2017)	*H. glycines*	13
(Natukunda et al., 2019)	*A. glycines*	5
(Chang et al., 2016)	*H. glycines*	25
(Vuong et al., 2015)	*H. glycines*	16
(Zhao et al., 2017)	*H. glycines*	13
(Chang and Hartman, 2017)	*A. glycines*	1
(Tran et al., 2019)	*H. glycines*	12
(Bao et al., 2014)	*H. glycines*	6
(Hanson et al., 2018)	*A. glycines*	45

2.2 Prediction of variant effects on genes

Variant effect prediction was performed with the function *predictCoding()* from the R package VariantAnnotation (Obenchain et al., 2014). Genome sequences and transcript coordinates were downloaded from PLAZA 4.0 (Van Bel et al., 2018). Reference and alternate alleles were manually extracted from each GWAS publication. Variants with no information on reference and alternate alleles in the original publication were discarded from this analysis.

2.3 Transcriptome data and selection of guide genes

Gene expression estimates in transcripts per million mapped reads (TPM, Kallisto estimation) were retrieved from the Soybean Expression Atlas (Machado et al., 2020). Additional RNA-seq samples comprising soybean tissues infested with pests were retrieved from a recent publication from our group (Almeida-Silva and Venancio, 2022). We filtered the GWAS and transcriptome datasets to keep only insect and nematode species that were represented by both data sources. We selected a total of 102 and 36 RNA-seq samples from soybean tissues infested with insects and nematodes, respectively (Supplementary Table S2). Finally, genes with median expression values lower than 5 were excluded to attenuate noise, resulting in a 15684 *x* 102 gene expression matrix for insects, and a 10240 x 36 matrix for nematodes. Guide genes were obtained from the Supplementary Data in (Almeida-Silva and Venancio, 2021c).

2.4 Candidate gene mining and functional analyses

Gene expression data were adjusted for confounding artifacts and quantile normalized with the R package BioNERO (Almeida-Silva and Venancio, 2021a). An unsigned coexpression network was inferred with BioNERO using Pearson's r as correlation. Candidate genes were identified and prioritized using the R package cageminer (Almeida-Silva and Venancio, 2021b) with default parameters. Module enrichment analyses were performed with BioNERO, using functional annotations from the PLAZA 4.0 database (Van Bel et al., 2018). Finally, prioritized candidates were given scores and ranks using the function *score_genes()* from cageminer.

2.5 Selection of most resistant accessions from the USDA germplasm

The VCF file with genotypic information for all accessions in the USDA germplasm was downloaded from Soybase (Brown et al., 2020). For each locus *i*, scores S_i 0, 1, or 2 were given based on the number of resistance-related SNPs. Total resistance scores for each accession were calculated as the sum of scores S_i for all *n* loci as follows:

$$S_{total} = \sum_{i=1}^{n} S_i \,, where\ S_i = \{0,1,2\}$$

Total resistance scores were ranked from highest to lowest, and ranks were used to select the most resistant accessions. The resistance potential of the best accessions was calculated as a ratio of the attributed scores to the theoretical maximum score (*2n,* which corresponds to all loci having scores 2).

3 Results and discussion

3.1 Data summary and genomic distribution of SNPs

After removing pest species that were not represented by both soybean RNA-seq and GWAS, our list of target species included the insects *A. glycines* (soybean aphid) and *S. litura* (armyworm caterpillar), and the nematode *H. glycines* (soybean cyst nematode) (Figure 1A). SNPs associated with resistance to all pest species were located in gene-rich regions of the soybean genome (Figure 1B), and their distributions were biased towards particular chromosomes (Figure 1C). Resistance SNPs against *A. glycines* were mostly located on chromosome 13, and resistance SNPs against *H. glycines* were mostly located on chromosomes 18, 8 and 7 (Figure 1C). Resistance SNPs against *S. litura* only occurred on chromosomes 12, 7, 6 and 5, but it is important to mention the small number of resistance SNPs against this species as compared to the other ones.

Interestingly, although most resistance SNPs against all species were located in intergenic regions, a considerable fraction of them was located in exons, except for *S. litura* (Figure 1D). This is a dramatic difference from what we observed in our previous study on fungi resistance-related genes, where almost all SNPs were located in intergenic regions (Almeida-Silva and Venancio, 2021c). Hence, we predicted SNP effects on coding sequences to better understand the functional consequences of these SNPs. From all resistance SNPs against *A. glycines* in coding regions, 31% (*n*=5) led to nonsynonymous substitutions, while 69% (*n*=11) led to synonymous substitutions (Supplementary Table S3). This unexpected finding suggests that most SNPs in coding regions increase resistance to this pest species despite not altering the amino acid residue. However, from all resistance SNPs against *H. glycines* in coding regions, nonsynonymous substitutions prevailed as expected (63%, *n*=10), followed by synonymous (31%, *n*=5) and nonsense substitutions (6%, *n*=1).

Additionally, we explored the distribution of SNPs in introns to understand their functional impact. SNPs in splice sites (*i.e.,* ±2 nucleotides relative to the exon-intron junction) have been shown to influence exon configuration and alternative splicing (Woolfe et al., 2010). None of the SNPs in introns were located in splice sites, indicating that they do not affect splicing patterns directly. This finding, together with the higher abundance of *A. glycines* resistance-related SNPs leading to synonymous substitutions, suggest that some SNPs contribute to resistance in non-canonical ways. A possible mechanistic explanation lies on recent advances in chromosome conformation capture (3C)-based methods (Wang et al., 2021). We hypothesize that these SNPs contribute to increased resistance through long-range interactions between genomic elements that are linked to transcriptional regulation, such as interactions in chromatin loops.

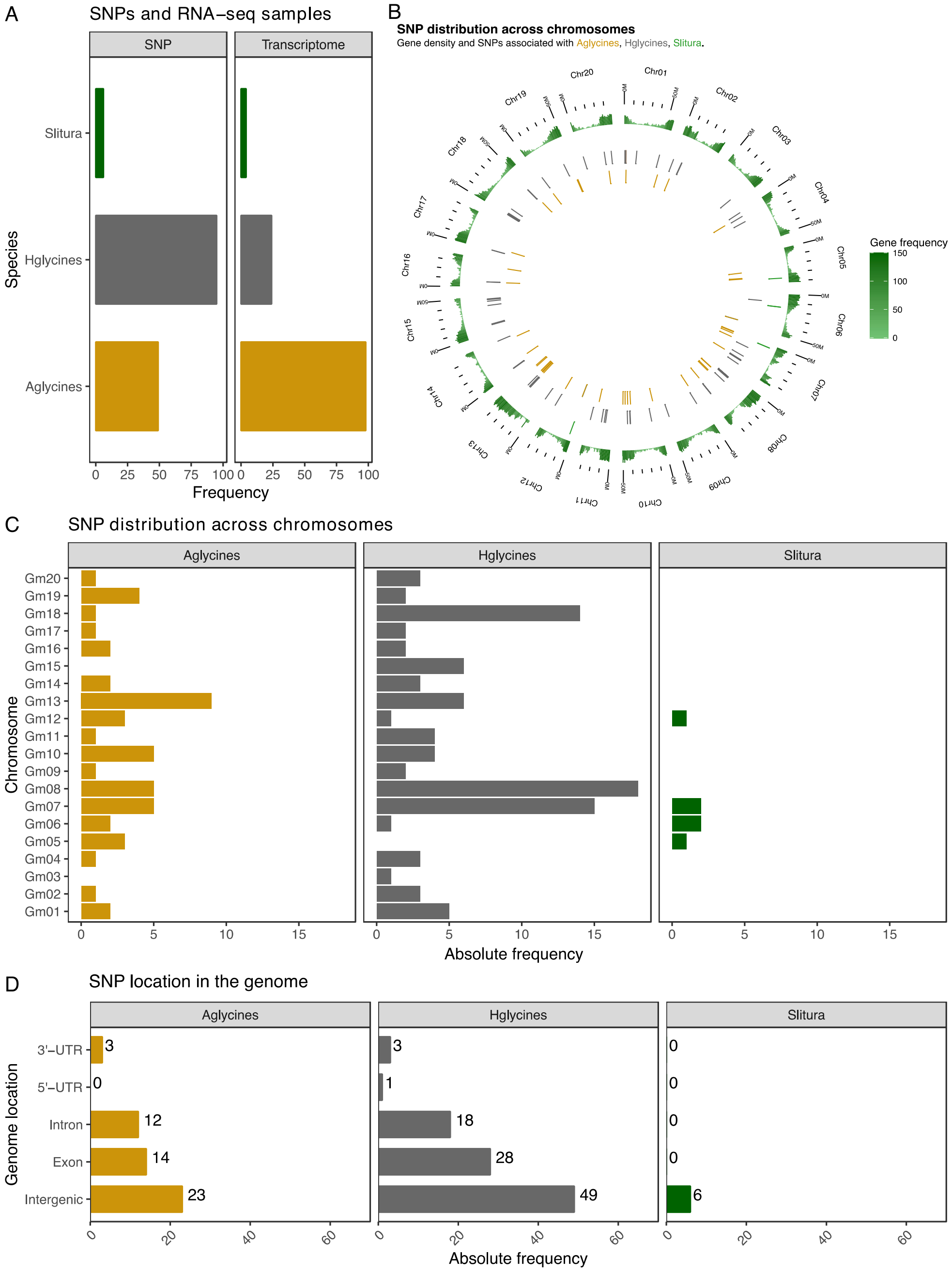


Figure 1. Data summary and genomic distribution of SNPs. A. Frequency of SNPs and RNA-seq samples included in this study. B. Genomic coordinates of resistance SNPs against each pest species. The outer track represents gene density, whereas inner tracks represent the SNP positions for each species. C. SNP distribution across chromosomes. Overall, there is an uneven distribution of SNPs across chromosomes. D. Genomic location of SNPs. Most SNPs are located in intergenic regions.

3.2 High candidate gene overlap between insects, but not between insects and nematodes

Using defense-related genes as guides, we identified 171, 7, and 228 high-confidence genes against *A. glycines*, *S. litura,* and *H. glycines*, respectively (Figure 2A)*.* Interestingly, 57% (4/7) of the candidates against *S. litura* were also candidates against *A. glycines*. However, none of the candidate resistance genes against insects were shared with *H. glycines*, revealing a high intraclass overlap (*i.e.,* among insects), but no interclass overlap (*i.e.,* among insects and nematodes). The shared genes are *Glyma.07G034400*, *Glyma.12G059900*, *Glyma.07G033100*, *Glyma.07G036400*, whose protein products are associated with phytohormone metabolism (KMD protein, Kelch repeat), transport (glucose and ATP transporters), and signaling (phospholipid:diacylglycerol acyltransferase), respectively. We also analyzed the overlap of pest resistance-related candidates with fungi resistance-related candidates from (Almeida-Silva and Venancio, 2021c) and found that a small number (n ≤5) of candidates against *H. glycines* and *A. glycines* are shared with *Fusarium* species (Figure 2B).

The observed overlap of candidate gene sets for different insect species is desirable, because it suggests that shared candidates can be used in biotechnological applications to equip soybean accessions with broad-spectrum resistance (BSR) against insects. In our recent study on candidate resistance genes against fungi, we reported a highly species-specific response (Almeida-Silva and Venancio, 2021c). This is an apparent trend for filamentous pathogens, as it has been reported in other studies (Ning and Wang, 2018; Kourelis and Van Der Hoorn, 2018). For insects, however, BSR to insects has been reported more often, and it can be achieved with genes associated with the synthesis of volatile organic compounds and secondary metabolites, for instance (Dixit et al., 2013; Vosman et al., 2018). Altogether, these findings suggest that achieving BSR against insects is easier than for filamentous pathogens, and it can be a feasible approach to control pests in soybean fields. This hypothesis can be tested in the future, when more data are available.

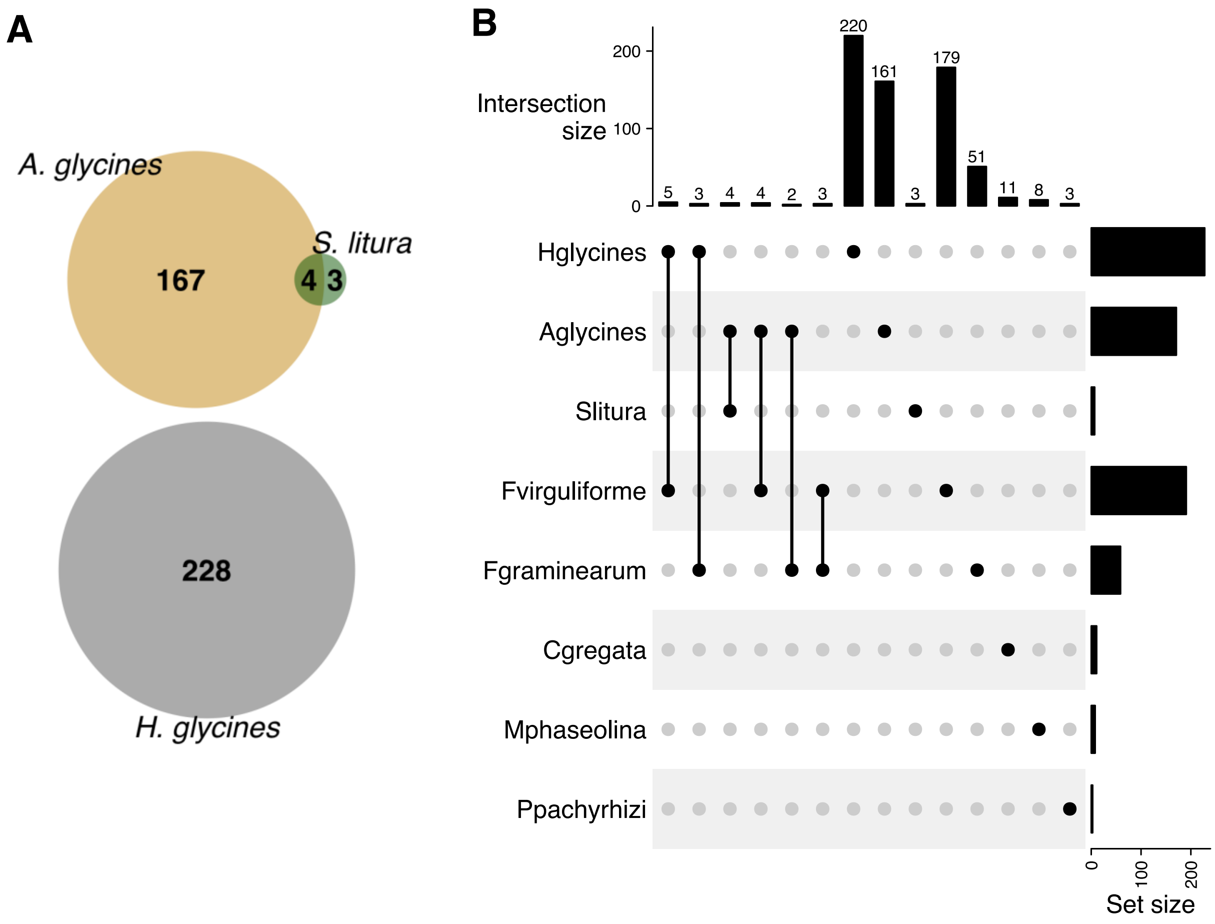


Figure 2. Cross-species overlap patterns across candidate gene sets. A. Euler diagram of prioritized candidate resistance genes against each pest species. Most candidate genes against *S. litura* are shared with *A. glycines*, suggesting a core defense against insects. However, insect resistance-related genes are not shared with nematode resistance-related genes. This suggests that insects and nematodes trigger different players of plant immunity. B. Upset plot with overlaps of candidate gene sets across fungal and pest species. A small number of candidate resistance genes against pests are shared with *Fusarium sp.* resistance-related gene sets. Candidate resistance genes against fungi were retrieved from (Almeida-Silva and Venancio, 2021c).

3.3 Signaling, oxidative stress and transcriptional regulation shape soybean resistance to pests

We manually curated the high-confidence candidate resistance genes to predict the putative role of their products in plant immunity (Supplementary Table S4). Most of the prioritized candidates encode proteins involved in immune signaling (23%), oxidative stress (21%), and transcriptional regulation (16%) (Figure 3). Candidates also encode proteins that play a role in transport, translational regulation, physical defense, phytohormone and secondary metabolism, apoptosis, recognition, and direct function against pests (Figure 3).

Interestingly, 55 (15%) candidate genes lack functional description and, hence, we could not infer their roles in resistance (*n*=28, 25, and 2 for *A. glycines, H. glycines,* and *S. litura*, respectively). Nevertheless, as they were identified as high-confidence candidate genes, we hypothesize that they encode defense-related proteins. This finding demonstrates that our algorithm can also serve as a network-based approach to predict functions of unannotated genes, similarly to previous approaches (Almeida-Silva et al., 2020; Depuydt and Vandepoele, 2021). Genes encoding proteins of unknown function were in the top 4 most abundant categories for all species, revealing a hidden, rich source of targets for biotechnological applications that would not have been identified if traditional SNP-to-gene mapping approaches were used.

We also developed a scheme that was used to rank high-confidence candidate genes (Table 2). As there are several candidate resistance genes against *A. glycines* and *H. glycines*, ranking candidates can help prioritize genes for validation purposes. Here, we suggest using the top 10 candidate resistance genes against each pathogen for experimental validation in future studies. Experimental tests with transgenic or edited soybeans using our set of target genes will likely reveal the most suitable candidates to develop soybean lines with increased resistance to each pest.

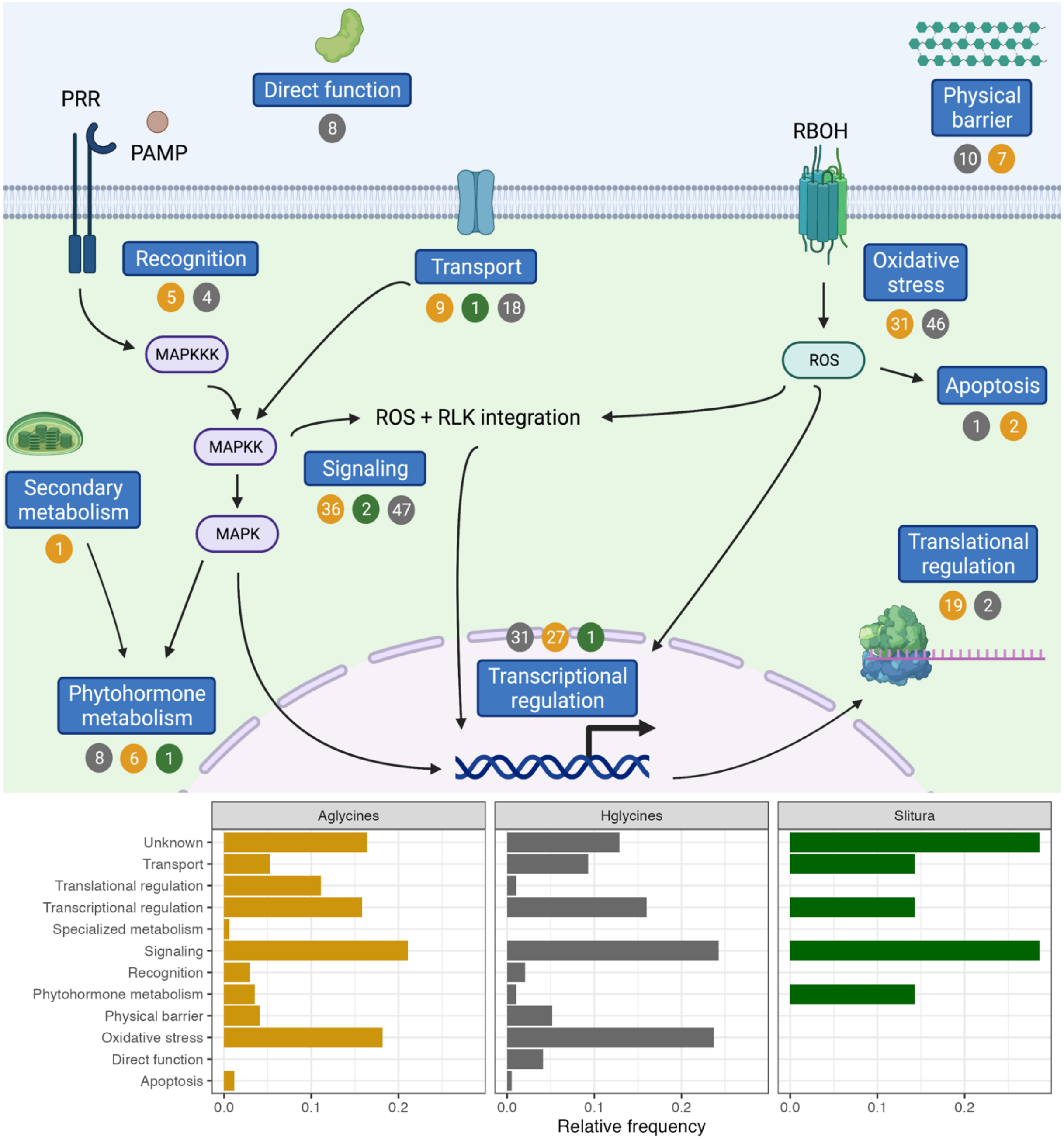


Figure 3. Prioritized candidate resistance genes and their putative role in plant immunity. Numbers in circles represent absolute frequencies of resistance genes against *Aphis glycines* (gold), *Heterodera glycines* (gray), and *Spodoptera litura* (green). PRR, pattern recognition receptor. PAMP, pathogen-associated molecular pattern. MAPKKK, mitogen-activated protein kinase kinase kinase. MAPKK, mitogen-activated protein kinase kinase. MAPK, mitogen-activated protein kinase. SAR, systemic acquired resistance. RBOH, respiratory burst oxidase homolog. ROS, reactive oxygen species. RLK, receptor-like kinase. Figure designed with Biorender (biorender.com).

Table 2. Top 10 candidate resistance genes against each pest species and their putative roles in plant immunity. The predicted function for each gene was manually curated from the description of the best ortholog in *Arabidopsis thaliana*, using functional annotations from Soybase and TAIR.

Gene	Predicted function	Resistance to	Role
Glyma.07G021800	Chaperone, DnaJ domain	*A. glycines*	Translational regulation
Glyma.19G187200	UDP-glycosyltransferase	*A. glycines*	Signaling
Glyma.10G002200	Calmodulin	*A. glycines*	Signaling
Glyma.11G194500	Acyl-CoA synthetase	*A. glycines*	Phytohormone metabolism
Glyma.01G210200	Autophagy protein	*A. glycines*	Apoptosis
Glyma.04G085500	Lysophospholipase	*A. glycines*	Signaling
Glyma.01G238800	Sugar transporter	*A. glycines*	Transport
Glyma.19G018600	AAA-ATPase	*A. glycines*	Oxidative stress
Glyma.13G326800	Galactose oxidase	*A. glycines*	Physical barrier
Glyma.04G085600	Unknown orphan gene	*A. glycines*	Unknown
Glyma.06G195300	Protein of unknown function	*S. litura*	Unknown
Glyma.07G033100	ATP transporter	*S. litura*	Transport
Glyma.07G034400	KMD protein - Kelch repeat	*S. litura*	Phytohormone metabolism
Glyma.06G175400	RNA-binding protein	*S. litura*	Transcriptional regulation
Glyma.05G194700	Protein of unknown function	*S. litura*	Unknown
Glyma.07G036400	Phospholipid:diacylglycerol acyltransferase	*S. litura*	Signaling
Glyma.12G059900	Glucose transporter	*S. litura*	Transport
Glyma.18G077500	Y-box transcription factor	*H. glycines*	Transcriptional regulation
Glyma.19G122700	MYB transcription factor	*H. glycines*	Transcriptional regulation
Glyma.13G169900	HD-Zip transcription factor	*H. glycines*	Transcriptional regulation
Glyma.09G245300	MYB transcription factor	*H. glycines*	Transcriptional regulation
Glyma.01G179900	Homeobox transcription factor	*H. glycines*	Transcriptional regulation
Glyma.16G053900	TCP transcription factor	*H. glycines*	Transcriptional regulation
Glyma.01G207300	HD-Zip transcription factor	*H. glycines*	Transcriptional regulation

3.4 Pangenome presence/absence variation analysis demonstrates that most prioritized genes are core genes

We analyzed PAV patterns for our prioritized candidate genes in the recently published pangenome of cultivated soybeans to unveil which soybean genotypes contain prioritized candidate genes and explore gene presence/absence variation patterns across genomes (Torkamaneh et al., 2021). We found that most candidates (98%) are present in all 204 accessions (Supplementary Figure 1A), similarly to what we found for fungi resistance-related genes (Almeida-Silva and Venancio, 2021c). This unsurprising trend is likely due to the high level of gene content conservation in this pangenome, which has 91% of the genes shared by >99% of the genomes.

Further, we investigated if gene PAV patterns could be explained by the geographical origins of the accessions (Supplementary Figure 1B). As we observed in our previous study (Almeida-Silva and Venancio, 2021c), PAV patterns did not cluster by geographical origin, suggesting that gene PAV is not affected by population structure. As this pangenome comprises improved soybean accessions (Torkamaneh et al., 2021), the lack of population structure effect can be due to breeding programs targeting optimal adaptation to different environmental conditions (*e.g.,* latitude and climate), even if they are in the same country.

3.5 Screening of the USDA germplasm reveals a room for genetic improvement

We inspected the USDA germplasm to find the top 10 most resistant genotypes against each pest species (see Materials and Methods for details). Strikingly, the most resistant genotypes do not contain all resistance alleles, revealing that, theoretically, they could be further improved to increase resistance (Table 3). None of the reported SNPs for resistance against *S. litura* have been identified in the SoySNP50k collection. Hence, we could not predict the most resistant accessions to this pest species in the USDA germplasm.

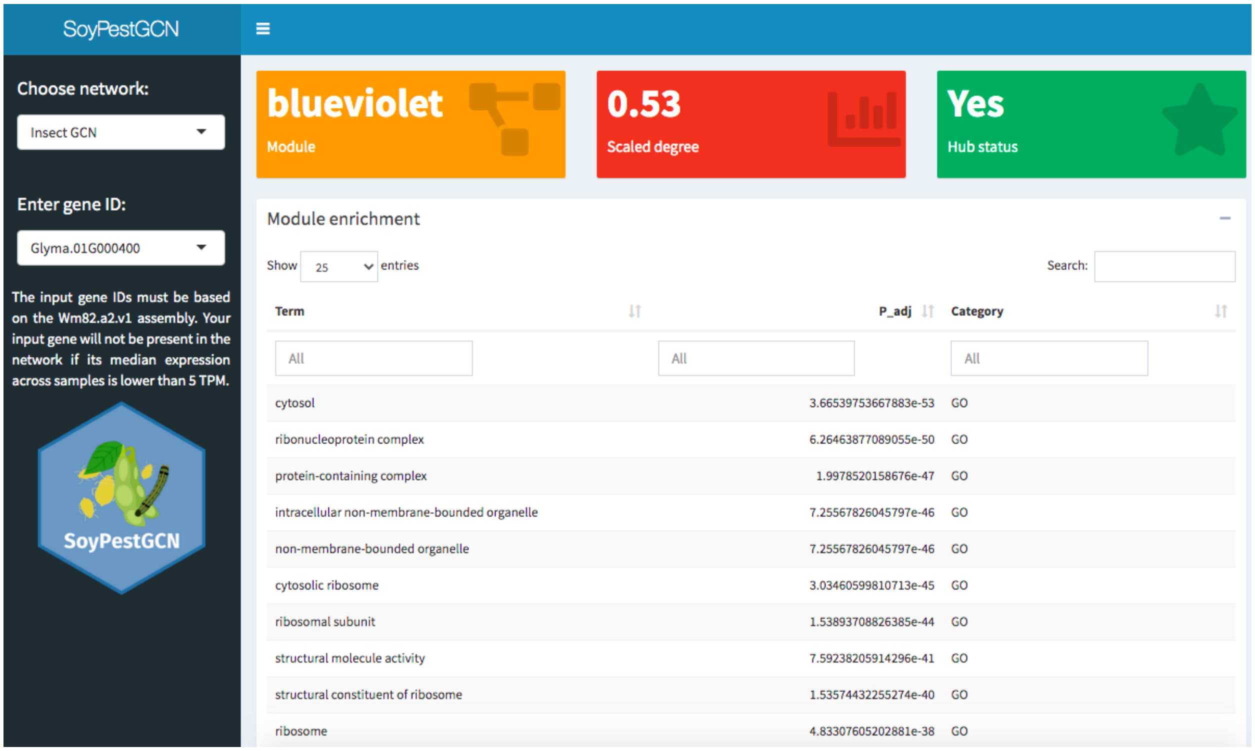
Our findings are in line with what we observed for resistance to fungi in the USDA germplasm (Almeida-Silva and Venancio, 2021c). Importantly, insect resistance potentials were lower than fungi resistance potentials (Wilcoxon rank-sum test, *P* = 7.7e-04), suggesting that pest resistance in could be further improved. A feasible approach to increase pest resistance involves pyramiding quantitative trait loci (QTL) that confer partial resistance to each pest (Li et al., 2020). To accomplish this, the most resistant genotypes identified here can be used in breeding programs using marker-assisted selection or inspire CRISPR/Cas editing strategies to introduce beneficial alleles or remove deleterious alleles, leading to increased resistance. However, as we are using independently published data, our model does not account for epistasis and different effect sizes for each variant. Hence, there might be accessions with a smaller number of SNPs with large effects that are more resistant than accessions with a greater number of SNPs with moderate effects.

Table 3. Top 10 most resistant soybean accessions against each pest species. Overall, the best genotypes do not reach the maximum potential. None of the resistance SNPs for *S. litura* have been identified in the USDA SoySNP50k compendium and, hence, we could not predict resistance potential against this species.

Accession	Score	Potential	Species
PI468399A	57	0.582	*A. glycines*
PI532451	57	0.582	*A. glycines*
PI468916	56	0.571	*A. glycines*
PI468918	56	0.571	*A. glycines*
PI479750	56	0.571	*A. glycines*
PI479752	56	0.571	*A. glycines*
PI468400B	55	0.561	*A. glycines*
PI468399B	54	0.551	*A. glycines*
PI507793	54	0.551	*A. glycines*
PI479749	54	0.551	*A. glycines*
PI556949	66	0.673	*H. glycines*
PI84751	66	0.673	*H. glycines*
Peking	66	0.673	*H. glycines*
PI438497	66	0.673	*H. glycines*
PI548402	66	0.673	*H. glycines*
PI438342	64	0.653	*H. glycines*
PI549047	64	0.653	*H. glycines*
PI597461C	64	0.653	*H. glycines*
PI404166	64	0.653	*H. glycines*
PI437679	64	0.653	*H. glycines*

3.6 Development of a user-friendly web application for network exploration

To facilitate network exploration and data reuse, we developed a user-friendly web application named SoyPestGCN (https://soypestgcn.venanciogroup.uenf.br/). Users can choose either the insect or the nematode GCN and input a soybean gene of interest (Wm82.a2.v1 assembly) to visualize the gene's module, scaled intramodular degree, and hub status (Figure 4A). Additionally, users can explore enriched GO terms, Mapman bins and/or Interpro domains associated with the input gene's module (Figure 4). This resource can be particularly useful for researchers studying soybean response to other pest species, as they can check if their genes of interest are located in defense-related coexpression modules. Also, researchers studying other species can verify if the soybean ortholog of their genes of interest is located in a defense-related module. The application is also available as an R package named SoyPestGCN (https://github.com/almeidasilvaf/SoyPestGCN). This package lets users run the application locally as a Shiny app, ensuring the application will always be available, even in case of server downtime.



Choose network: Insect GCN

Enter gene ID: Glyma.01G000400

The input gene IDs must be based on the Wm82.a2.v1 assembly. Your input gene will not be present in the network if its median expression across samples is lower than 5 TPM.

blueviolet
Module

0.53
Scaled degree

Yes
Hub status

Module enrichment

Show 25 entries Search:

Term	P_adj	Category
All	All	All
cytosol	3.66539753667883e-53	GO
ribonucleoprotein complex	6.26463877089055e-50	GO
protein-containing complex	1.9978520158676e-47	GO
intracellular non-membrane-bounded organelle	7.25567826045797e-46	GO
non-membrane-bounded organelle	7.25567826045797e-46	GO
cytosolic ribosome	3.03460599810713e-45	GO
ribosomal subunit	1.53893708826385e-44	GO
structural molecule activity	7.59238205914296e-41	GO
structural constituent of ribosome	1.53574432255274e-40	GO
ribosome	4.83307605202881e-38	GO

Figure 4. Functionalities in the SoyPestGCN web application. Screenshot of the page users see when they access the application. In the sidebar, users can specify either the insect GCN or the nematode GCN followed by the ID of a gene of interest (Wm82.a2.v1 assembly). For each gene, users can see the gene's module (orange box), scaled degree (red box), hub gene status (green box), and an interactive table with enrichment results for MapMan bins, Interpro domains and Gene Ontology terms associated the gene's module. P-values from enrichment results are adjusted for multiple testing with Benjamini-Hochberg correction.

4 Conclusions

By integrating publicly available GWAS and RNA-seq data, we found promising candidate genes in soybean associated with resistance to three pest species, namely *A. glycines, S. litura*, and *H. glycines.* The prioritized candidates encode proteins that play a role immunity-related processes such as in recognition, signaling, transcriptional regulation, oxidative stress, specialized metabolism, and physical defense. We have also found the top 10 most resistant soybean accessions against each pest species and hypothesize that they can be used in soybean improvement programs, either via breeding with marker-assisted selection or through genome editing. The coexpression network generated here was also made available in a web resource and R package to help in future studies on soybean-pest interactions.

Data availability

All data and code used in this study are available in our GitHub repository (https://github.com/almeidasilvaf/SoyPestGCN_paper) to ensure full reproducibility.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.

Author contributions

Conceived the study: FA-S and TMV. Data analysis: FA-S. Funding, project coordination and infrastructure: TMV. Manuscript writing: FA-S and TMV.

Conflicts of interest: none.

REFERENCES

Almeida-Silva, F., Moharana, K.C., Machado, F.B., and Venancio, T.M. (2020). Exploring the complexity of soybean (Glycine max) transcriptional regulation using global gene co-expression networks. Planta **252**: 1–12.

Almeida-Silva, F. and Venancio, T.M. (2021a). BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction. Funct. Integr. Genomics.

Almeida-Silva, F. and Venancio, T.M. (2021b). cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks. bioRxiv: 1–9.

Almeida-Silva, F. and Venancio, T.M. (2021c). Integration of genome-wide association studies and gene coexpression networks unveils promising soybean resistance genes against five common fungal pathogens. bioRxiv: 1–19.

Almeida-Silva, F. and Venancio, T.M. (2022). Pathogenesis-related protein 1 (PR-1) genes in soybean: Genome-wide identification, structural analysis and expression profiling under multiple biotic and abiotic stresses. Gene **809**: 146013.

Bandara, A.Y., Weerasooriya, D.K., Bradley, C.A., Allen, T.W., and Esker, P.D. (2020). Dissecting the economic impact of soybean diseases in the United States over two decades. PLoS One **15**: 1–28.

Bao, Y., Vuong, T., Meinhardt, C., Tiffin, P., Denny, R., Chen, S., Nguyen, H.T., Orf, J.H., and Young, N.D. (2014). Potential of Association Mapping and Genomic Selection to Explore PI 88788 Derived Soybean Cyst Nematode Resistance. Plant Genome **7**: plantgenome2013.11.0039.

Baxter, I. (2020). We aren't good at picking candidate genes, and it's slowing us down. Curr. Opin. Plant Biol. **54**: 57–60.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van De Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46**: D1190–D1196.

Brown, A. V, Conners, S.I., Huang, W., Wilkey, A.P., Grant, D., Weeks, N.T., Cannon, S.B., Graham, M.A., and Nelson, R.T. (2020). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. **13**: 1–6.

Chang, H.X. and Hartman, G.L. (2017). Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Front. Plant Sci. **8**: 1–12.

Chang, H.X., Lipka, A.E., Domier, L.L., and Hartman, G.L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Phytopathology **106**: 1139–1151.

Depuydt, T. and Vandepoele, K. (2021). Multi-omics network-based functional annotation of unknown Arabidopsis genes. Plant J. **n/a**.

Dixit, S., Upadhyay, S.K., Singh, H., Sidhu, O.P., Verma, P.C., and Chandrashekar, K. (2013). Enhanced methanol production in plants provides broad spectrum insect resistance. PLoS One **8**: 1–13.

Hanson, A.A., Lorenz, A.J., Hesler, L.S., Bhusal, S.J., Bansal, R., Michel, A.P., Jiang, G., and Koch, R.L. (2018). Genome-Wide Association Mapping of Host-Plant Resistance to Soybean Aphid. Plant Genome **11**: 1–12.

Koenning, S.R. and Wrather, J.A. (2010). Suppression of Soybean Yield Potential in the Continental United States by Plant Diseases from 2006 to 2009. Plant Heal. Prog. **11**: 5.

Kourelis, J. and Van Der Hoorn, R.A.L. (2018). Defended to the Nines: 25 years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. Plant Cell.

Li, W., Deng, Y., Ning, Y., He, Z., and Wang, G.L. (2020). Exploiting Broad-Spectrum Disease Resistance in Crops: From Molecular Dissection to Breeding. Annu. Rev. Plant Biol. **71**: 575–603.

Liu, H., Che, Z., Zeng, X., Zhang, G., Wang, H., and Yu, D. (2016). Identification of single nucleotide polymorphisms in soybean associated with resistance to common cutworm (Spodoptera litura Fabricius). Euphytica **209**: 49–62.

Liu, S., Ge, F., Huang, W., Lightfoot, D.A., and Peng, D. (2019). Effective identification of soybean candidate genes involved in resistance to soybean cyst nematode via direct whole genome re-sequencing of two segregating mutants. Theor. Appl. Genet. **132**: 2677–2687.

Machado, F.B., Moharana, K.C., Almeida-Silva, F., Gazara, R.K., Pedrosa-Silva, F., Coelho, F.S., Grativol, C., and Venancio, T.M. (2020). Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas. Plant J.: 0–2.

Michno, J.M., Liu, J., Jeffers, J.R., Stupar, R.M., and Myers, C.L. (2020). Identification of nodulation-related genes in Medicago truncatula using genome-wide association studies and co-expression networks. Plant Direct **4**: 1–10.

Natukunda, M.I., Parmley, K.A., Hohenstein, J.D., Assefa, T., Zhang, J., Macintosh, G.C., and Singh, A.K. (2019). Identification and Genetic Characterization of Soybean Accessions Exhibiting Antibiosis and Antixenosis Resistance to Aphis glycines (Hemiptera: Aphididae). J. Econ. Entomol. **112**: 1428–1438.

Ning, Y. and Wang, G.L. (2018). Breeding plant broad-spectrum resistance without yield penalties. Proc. Natl. Acad. Sci. U. S. A. **115**: 2859–2861.

Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. Bioinformatics **30**: 2076–2078.

Oliveira, C.M., Auad, A.M., Mendes, S.M., and Frizzas, M.R. (2014). Crop losses and the economic impact of insect pests on Brazilian agriculture. Crop Prot. **56**: 50–54.

Torkamaneh, D., Lemay, M.-A., and Belzile, F. (2021). The Pan-genome of the Cultivated Soybean (PanSoy) Reveals an Extraordinarily Conserved Gene Content. Plant Biotechnol. J. **n/a**.

Tran, D.T., Steketee, C.J., Boehm, J.D., Noe, J., and Li, Z. (2019). Genome-wide association analysis pinpoints additional major genomic regions conferring resistance to soybean cyst nematode (Heterodera glycines ichinohe). Front. Plant Sci. **10**: 1–13.

Vosman, B., van't Westende, W.P.C., Henken, B., van Eekelen, H.D.L.M., de Vos, R.C.H., and Voorrips, R.E. (2018). Broad spectrum insect resistance and metabolites in close relatives of the cultivated tomato. Euphytica **214**.

Vuong, T.D., Sonah, H., Meinhardt, C.G., Deshmukh, R., Kadam, S., Nelson, R.L., Shannon, J.G., and Nguyen, H.T. (2015). Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC Genomics **16**: 1–13.

Wang, L., Jia, G., Jiang, X., Cao, S., Chen, Z.J., and Song, Q. (2021). Altered chromatin architecture and gene expression during polyploidization and domestication of soybean. Plant Cell: 1430–1446.

Woolfe, A., Mullikin, J.C., and Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. Genome Biol. **11**.

Zhang, J., Wen, Z., Li, W., Zhang, Y., Zhang, L., Dai, H., Wang, D., and Xu, R. (2017). Genome-wide association study for soybean cyst nematode resistance in Chinese elite soybean cultivars. Mol. Breed. **37**.

Zhao, X., Teng, W., Li, Y., Liu, D., Cao, G., Li, D., Qiu, L., Zheng, H., Han, Y., and Li, W. (2017). Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5.7. BMC Genomics **18**: 1–10.

Supplementary Figures

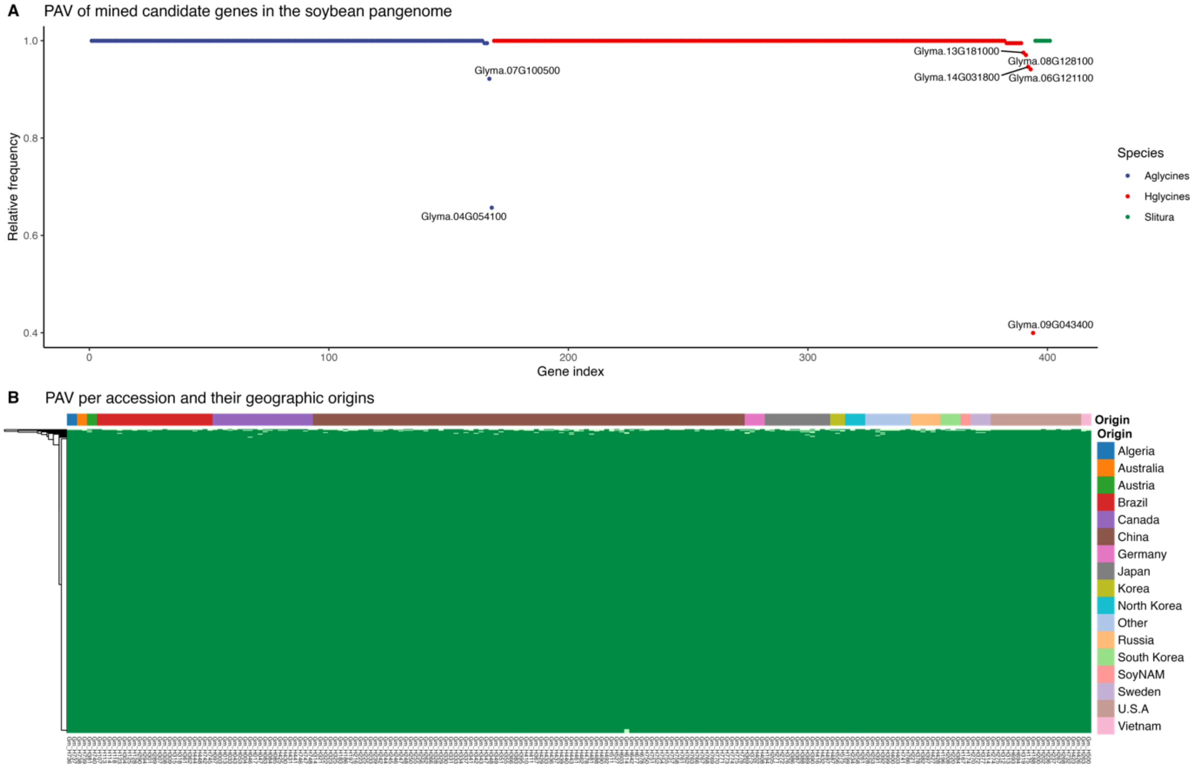


Figure 1. Presence/absence variation (PAV) of prioritized candidate genes in the soybean pangenome. A. Relative frequency of accessions containing each candidate gene. Most candidates are present in all accessions. Candidate genes with lower frequency in the pangenome are labeled. B. PAV per accessions and their geographic distribution. The patterns of gene PAV cannot be explained by the geographic origins of the accessions.

Other publications

The state of the art in soybean transcriptomics resources and gene coexpression networks

Fabricio Almeida-Silva[1], Kanhu C. Moharana[1], Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Review

Situation: Published (DOI: 10.1093/insilicoplants/diab005)

Journal: *in silico* Plants

Impact factor: no; CiteScore: 2.8; Highest percentile: 67%

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

ABSTRACT
In the past decade, over 3000 samples of soybean transcriptomic data have accumulated in public repositories. Here, we review the state of the art in soybean transcriptomics, highlighting the major microarray and RNA-seq studies that investigated soybean transcriptional programs in different tissues and conditions. Further, we propose approaches for integrating such big data using gene coexpression network and outline important web resources that may facilitate soybean data acquisition and analysis, contributing to the acceleration of soybean breeding and functional genomics research.

Keywords: legumes, functional genomics, gene expression, hub genes, modules.

Pathogenesis-related protein 1 (PR-1) genes in soybean: genome-wide identification, structural analysis and expression profiling under multiple biotic and abiotic stresses

Fabricio Almeida-Silva[1], Thiago M. Venancio[1*]

[1]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

Type of article: Research article

Situation: Published (DOI: 10.1016/j.gene.2021.146013)

Journal: Gene

Impact factor: 3.688; CiteScore: 5.6; Highest percentile: 66%

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

ABSTRACT

Plant pathogenesis-related (PR) proteins are a large group of proteins, classified in 17 families, that are induced by pathological conditions. Here, we characterized the soybean PR-1 (GmPR-1) gene repertoire at the sequence, structural and expression levels. We found 24 GmPR-1 genes, clustered in two phylogenetic groups. GmPR-1 genes are under strong purifying selection, particularly those that emerged by tandem duplications. GmPR-1 promoter regions are abundant in cis-regulatory elements associated with major stress-related transcription factor families, namely WRKY, ERF, HD-Zip, C2H2, NAC, and GATA. We observed that 23 GmPR-1 genes are induced by stress conditions or exclusively expressed upon stress. We explored 1972 transcriptome samples, including 26 stress conditions, revealing that most GmPR-1 genes are differentially expressed in a plethora of biotic and abiotic stresses. Our findings highlight stress-responsive GmPR-1 genes with potential biotechnological applications, such as the development of transgenic lines with increased resistance to biotic and abiotic stresses.

CONCLUSÕES GERAIS

Conclusões Gerais

Neste trabalho, desenvolvemos um pacote R chamado *BioNERO* destinado a facilitar a inferência e análise de redes biológicas (*i.e.,* redes de coexpressão, redes regulatórias, e redes de interação proteína-proteína). Além disso, *BioNERO* permite a comparação de redes inferidas a partir de dados transcriptômicos obtidos de indivíduos da mesma espécie e de espécies diferentes, podendo ser usado para análises evolutivas.

Além disso, desenvolvemos um pacote R chamado *cageminer* destinado a integrar redes de coexpressão gênica inferidas com o *BioNERO* com SNPs obtidos de GWAS para identificar e priorizar genes candidatos de alta confiança associados a características quantitativas. O algoritmo implementado no *cageminer* reduz as listas de possíveis genes candidatos em 99%, resultando em pequenos conjuntos de genes de alta confiança que são potenciais alvos para aplicações biotecnológicas.

Finalmente, aplicamos os dois pacotes desenvolvidos nesse trabalho para identificar e priorizar genes candidatos em soja associados à resistência a: i. fungos fitopatogênicos, em particular *Cadophora gregata, Fusarium virguliforme, Fusarium graminearum, Macrophomina phaseolina*, e *Phakopsora pachyrhizi* e; ii. pragas da cultura da soja, em particular os insetos *Aphis glycines* e *Spodoptera litura*, e o nematoide *Heterodera glycines.* Os genes identificados em cada estudo devem ser validados experimentalmente e podem ser usados para desenvolver linhagens de soja resistentes a fungos e pragas.