



INVESTIGAÇÃO GENÔMICA DE ESPÉCIES SILVESTRES DO GÊNERO *GLYCINE*
PARA ELUCIDAR A EVOLUÇÃO E DIVERSIDADE DE GENES COM RELEVÂNCIA
AGRONÔMICA EM SOJA

DIEGO DAVID ASTETE PEÑA

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO - UENF

CAMPOS DOS GOYTACAZES - RJ

FEVEREIRO - 2026

INVESTIGAÇÃO GENÔMICA DE ESPÉCIES SILVESTRES DO GÊNERO *GLYCINE*
PARA ELUCIDAR A EVOLUÇÃO E DIVERSIDADE DE GENES COM RELEVÂNCIA
AGRONÔMICA EM SOJA

DIEGO DAVID ASTETE PEÑA

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

Orientador:

Dr. Thiago Motta Venancio.

Coorientador:

Dr. Gabriel Quintanilha Peixoto

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO - UENF

CAMPOS DOS GOYTACAZES - RJ

FEVEREIRO - 2026

FICHA CATALOGRÁFICA

UENF - Bibliotecas

Elaborada com os dados fornecidos pelo autor.

P397

Peña, Diego David Astete.

Investigação genômica de espécies silvestres do gênero *Glycine* para elucidar a evolução e diversidade de genes com relevância agrônômica em soja. / Diego David Astete Peña. - Campos dos Goytacazes, RJ, 2026.

72 f. : il.

Inclui bibliografia.

Dissertação (Mestrado em Biotecnologia Vegetal) - Universidade Estadual do Norte Fluminense Darcy Ribeiro, Centro de Biociências e Biotecnologia, 2026.

Orientador: Thiago Motta Venancio.

Coorientador: Gabriel Quintanilha Peixoto.

1. pangenoma. 2. expressão gênica. 3. conservação gênica. 4. transcriptômico. I. Universidade Estadual do Norte Fluminense Darcy Ribeiro. II. Título.

CDD - 660.6

INVESTIGAÇÃO GENÔMICA DE ESPÉCIES SILVESTRES DO GÊNERO *GLYCINE*
PARA ELUCIDAR A EVOLUÇÃO E DIVERSIDADE DE GENES COM RELEVÂNCIA
AGRONÔMICA EM SOJA

DIEGO DAVID ASTETE PEÑA

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

Aprovado em 24 de fevereiro de 2026

Comissão Examinadora:



Documento assinado digitalmente
ANTONIA ELENIR AMANCIO OLIVEIRA
Data: 16/04/2026 11:34:33-0300
Verifique em <https://validar.iti.gov.br>

Dra. Antônia Elenir Amâncio Oliveira (Doutora em Biociências e Biotecnologia) -
UENF



Documento assinado digitalmente
RODRIGO NUNES DA FONSECA
Data: 24/04/2026 09:54:49-0300
Verifique em <https://validar.iti.gov.br>

Dr. Rodrigo Nunes da Fonseca (Doutor em Genética e Genômica Funcional) - UFRJ



Documento assinado digitalmente
SARAH RODRIGUES FERREIRA
Data: 16/04/2026 12:37:31-0300
Verifique em <https://validar.iti.gov.br>

Dra. Sarah Rodrigues Ferreira (Doutora em Biotecnologia Vegetal) - UENF



Documento assinado digitalmente
GABRIEL QUINTANILHA PEIXOTO
Data: 17/04/2026 10:18:27-0300
Verifique em <https://validar.iti.gov.br>

Dr. Gabriel Quintanilha Peixoto (Doutor em Bioinformática) - UENF

Coorientador



Documento assinado digitalmente
THIAGO MOTTA VENANCIO
Data: 16/04/2026 14:17:53-0300
Verifique em <https://validar.iti.gov.br>

Dr. Thiago Motta Venancio (Doutor em Bioinformática) - UENF
Orientador

DEDICATÓRIA

À memória do meu querido avô Roger,
que sempre acreditou em mim e que, em vida, foi
meu companheiro de sonhos e risadas. Ele viverá
para sempre no meu coração.

AGRADECIMENTOS

Primeiramente rendo graças a Deus, por me conceder saúde, força e perseverança ao longo de toda essa jornada. Nos momentos de incerteza e cansaço, foi na fé que encontrei equilíbrio, esperança e motivação para seguir em frente.

À minha família, ao meu pai Helvio, à minha mãe Monica, à minha irmã Kendall e à minha tia Amy, que sempre estiveram ao meu lado, oferecendo apoio incondicional, compreensão e incentivo em todos os momentos. Nada disso teria sido possível sem o amor, a paciência e os valores que me ensinaram desde sempre.

À minha namorada Maria, pelo carinho, companheirismo e apoio constante. Obrigado por estar presente nos momentos difíceis, por compreender as ausências e por acreditar em mim mesmo quando eu duvidei. Nossas noites de filmes e jogos foram essenciais para me sentir ainda mais próximo de você e para aliviar a mente nos momentos mais desafiadores dessa trajetória. Seu apoio foi fundamental para que eu chegasse até aqui.

Aos meus orientadores, Dr. Thiago Motta Venancio e Dr. Gabriel Quintanilha Peixoto, pela orientação científica, dedicação e confiança ao longo deste trabalho. Agradeço especialmente pela paciência, pelas discussões construtivas, pelo rigor acadêmico e por acreditarem em mim e no meu potencial em todos os momentos. Sei que o processo não foi simples, e justamente por isso valorizo ainda mais o apoio, a compreensão e o comprometimento demonstrados ao longo de toda essa trajetória, que foram fundamentais para o meu crescimento profissional e intelectual.

Aos amigos e colegas de laboratório, pelo convívio diário, pelas trocas de conhecimento, pelas discussões científicas e também pelos momentos de descontração que tornaram essa caminhada mais leve. Agradeço em especial ao meu amigo Daniel, pelo apoio constante, pela ajuda acadêmica e pelas contribuições fundamentais ao longo do desenvolvimento deste trabalho. O apoio coletivo foi essencial para enfrentar os desafios do mestrado.

A todos os amigos que, direta ou indiretamente, contribuíram para a realização deste trabalho, deixo meu sincero agradecimento. Em especial, a Rodrigo e Emmanuel, meus grandes amigos, pelo apoio emocional constante, pelas conversas, pelas noites de jogos e pelos momentos de descontração que tornaram essa jornada mais leve. A Hector, Remy, Yhony e Adan, com quem compartilhei a convivência ao longo de toda a pós-graduação, agradeço pelo companheirismo, pelas trocas diárias e pelo apoio ao longo de todo o mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Código de Financiamento 001), da Fundação Carlos Chagas de Amparo à Pesquisa no Estado do Rio de Janeiro (FAPERJ) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

LISTA DE FIGURAS

FIGURA 1: Distribuição global da produção de soja por país, expressa em toneladas métricas.

FIGURA 2. Relações filogenéticas, distribuição geográfica e estimativas de tempo de divergência entre espécies anuais e perenes do gênero *Glycine*.

FIGURA 3: Padrão de conservação de genes tecido-específicos de *G. max* em espécies perenes.

FIGURA 4: Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos vegetais considerando genes TS não conservados.

FIGURA 5: Análise de enriquecimento de termos Gene Ontology (GO) na raiz, considerando genes TS conservados.

FIGURA 6: Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos de *Glycine soja*, considerando genes não conservados.

FIGURA 7: Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos de *Glycine soja*, considerando genes conservados.

LISTA DE TABELAS

TABELA 1: Associação entre conservação gênica e especificidade tecidual em espécies silvestres do gênero *Glycine*.

TABELA 2: Associação entre conservação gênica e especificidade tecidual em *G. soja*.

TABELA 3: Informações sobre os genes mais representativos em todos os tecidos em espécies perenes.

TABELA 4: Informações sobre os genes mais representativos em todos os tecidos de *G. soja*.

TABELA S1: Termos de Gene Ontology (GO) enriquecidos para os genes tecido-específicos nas espécies perenes do gênero *Glycine*.

TABELA S2: Termos de Gene Ontology (GO) enriquecidos para os genes tecido-específicos em *G. soja*.

TABELA S3: Ortogrupos inspecionados manualmente para prospecção de genes de interesse.

LISTA DE ABREVIACÕES

ADP: Adenosina difosfato

BP: *Biological Process* (Processo Biológico)

CASP: *Casparian Strip Membrane Protein*

CRISPR: *Clustered Regularly Interspaced Short Palindromic Repeats*

FBN: Fixação Biológica de Nitrogênio

GO: *Gene Ontology*

GWAS: *Genome-Wide Association Studies*

MLP-like: *Major Latex Protein-like*

NB-ARC: *Nucleotide-Binding Apaf-1, R proteins and CED-4*

N-TS: Não Tecido-Específico

OR: *Odds Ratio*

RNA-Seq: *RNA Sequencing*

SEA: *Soybean Expression Atlas*

SNP: *Single Nucleotide Polymorphism*

TS: Tecido-Específico

WGD: *Whole Genome Duplication*

SUMÁRIO

RESUMO.....	11
ABSTRACT.....	12
INTRODUÇÃO.....	1
REVISÃO BIBLIOGRÁFICA.....	4
Diversidade genética e limitações da soja cultivada.....	4
O gênero <i>Glycine</i> : diversidade, divergência e potencial evolutivo.....	4
Domesticação da soja.....	7
Criação do primeiro pangenoma da soja e evolução do conteúdo gênico.....	9
Expressão gênica e especialização tecidual.....	10
OBJETIVO GERAL.....	12
OBJETIVOS ESPECÍFICOS.....	12
MÉTODOS.....	13
Bases de dados e conjunto genômico.....	13
Seleção e classificação dos genes.....	13
Construção da matriz binária.....	14
Análise estatística.....	15
Enriquecimento funcional.....	16
RESULTADOS.....	17
Análise funcional e enriquecimento de genes tecido-específicos.....	21
Enriquecimento GO dos genes TS em <i>G. max</i> e perdidos em <i>Glycine soja</i>	30
Enriquecimento GO dos genes TS em <i>G. max</i> e conservados em <i>Glycine soja</i>	33
DISCUSSÃO.....	36
REFERÊNCIAS BIBLIOGRÁFICAS.....	41

RESUMO

O gênero *Glycine* inclui espécies anuais e perenes que representam uma fonte importante de diversidade genética para o melhoramento da soja cultivada. Neste estudo, investigamos a evolução e a conservação de genes tecido-específicos (TS) utilizando um super-pangenoma composto por 57 genomas de espécies perenes e anuais do gênero *Glycine*, integrado com dados de expressão gênica e análises funcionais. Padrões contrastantes foram identificados entre os tecidos em relação à conservação e expressão gênica: a maioria dos genes TS, especialmente aqueles relacionados ao embrião, à proteção da semente e a estruturas reprodutivas, tendem a não ser conservados em espécies perenes. Tais genes são enriquecidos em processos associados ao desenvolvimento, metabolismo energético, defesa química e sinalização hormonal, indicando forte diversificação funcional em espécies perenes. Em contraste, a raiz destacou-se como o único tecido com associação positiva entre conservação e expressão tecido-específica, indicando que tal padrão de expressão precede a diversificação do gênero *Glycine* nos ramos perene e anual. Complementarmente, análises independentes em *Glycine soja* permitiram comparar os padrões de conservação dos genes TS de *G. max* entre espécies perenes e a espécie anual silvestre. Os resultados indicam divergência funcional significativa, especialmente em tecidos associados ao desenvolvimento da semente, enquanto tecidos essenciais, como raiz, flor e nódulo, mantêm maior conservação gênica. Além disso, a identificação de ortogrupos exclusivos e genes candidatos nas espécies perenes, associados a processos de defesa, metabolismo e desenvolvimento, evidencia que esses genomas preservam variantes funcionais alternativas não presentes na soja cultivada. Esses achados destacam o papel das espécies perenes como importantes reservatórios de diversidade genética potencialmente explorável no melhoramento da soja. Este trabalho demonstra que a integração de dados pangenômicos e transcriptômicos fornece uma abordagem promissora para compreender a evolução funcional de genes de interesse biotecnológico no gênero *Glycine*.

Palavras-chave: pangenoma; expressão gênica; conservação gênica; transcriptômica.

ABSTRACT

The genus *Glycine* includes annual and perennial species that represent an important source of genetic diversity for cultivated soybean improvement. In this study, we investigated the evolution and conservation of tissue-specific (TS) genes using a super-pangenome comprising 57 genomes from perennial and annual species of the genus *Glycine*, integrated with gene expression data and functional analyses. Contrasting patterns were identified across tissues regarding gene conservation and expression: most TS genes, especially those related to the embryo, seed protection, and reproductive structures, tend not to be conserved in perennial species. Such genes are enriched in processes associated with development, energy metabolism, chemical defense, and hormonal signaling, indicating strong functional diversification in perennial species. In contrast, the root stood out as the only tissue with a positive association between conservation and tissue-specific expression, indicating that this expression pattern predates the diversification of the genus *Glycine* into perennial and annual lineages. Complementarily, independent analyses in *Glycine soja* allowed comparison of the conservation patterns of *G. max* TS genes between perennial species and the wild annual species. The results indicate significant functional divergence, particularly in tissues associated with seed development, while essential tissues such as root, flower, and nodule maintain greater gene conservation. Furthermore, the identification of exclusive orthogroups and candidate genes in perennial species, associated with defense, metabolism, and developmental processes, demonstrates that these genomes harbor alternative functional variants absent from cultivated soybean. These findings highlight the role of perennial species as important reservoirs of genetic diversity potentially exploitable in soybean breeding. This work demonstrates that the integration of pangenomic and transcriptomic data provides a promising approach for understanding the functional evolution of genes of biotechnological interest in the genus *Glycine*.

Keywords: pangenome; gene expression; gene conservation; transcriptomics.

INTRODUÇÃO

A soja (*Glycine max* [L.] Merr.) é uma das culturas agrícolas mais relevantes do mundo, desempenhando papel estratégico na segurança alimentar, na economia global e na sustentabilidade dos sistemas produtivos (Xu *et al.*, 2022). Com elevado teor de proteína e óleo, constitui a principal fonte de alimento para a indústria de rações, além de ser amplamente utilizada na alimentação humana e na produção de biocombustíveis (Graham e Vance, 2003; Lee *et al.*, 2011). Em 2022, Brasil, Estados Unidos e Argentina responderam por mais de 80% da produção global, reforçando a relevância econômica dessa cultura para o cenário agrícola internacional (Soystats, 2025). O óleo derivado dessa cultura representa cerca de 59% da produção mundial de oleaginosas (Abdelghany *et al.*, 2020).

A soja pertence à família Fabaceae, caracterizada pela capacidade de estabelecer associações simbióticas com bactérias diazotróficas responsáveis pela fixação biológica de nitrogênio (FBN) (Schmutz *et al.*, 2010). Tal característica reduz a dependência de fertilizantes nitrogenados e contribui para sistemas de produção mais sustentáveis. A soja também foi a primeira leguminosa a ter seu genoma completamente sequenciado, marco que impulsionou avanços significativos em genômica, transcriptômica e biotecnologia vegetal (Schmutz *et al.*, 2010). Apesar de sua relevância agrônômica, a soja cultivada apresenta baixa diversidade genética, consequência de um longo processo de domesticação, seleção e melhoramento. Essa redução de variabilidade limita as possibilidades de melhoramento convencional visando a adaptação às mudanças climáticas, à incidência de novos patógenos e às condições ambientais extremas, tornando urgente a investigação de novas fontes de diversidade genética a serem exploradas por meio da introdução de genes e alelos através de transformação ou edição genética (Razzaq *et al.*, 2021; Sedivy, Wu e Hanzawa, 2017; Zhuang *et al.*, 2022).

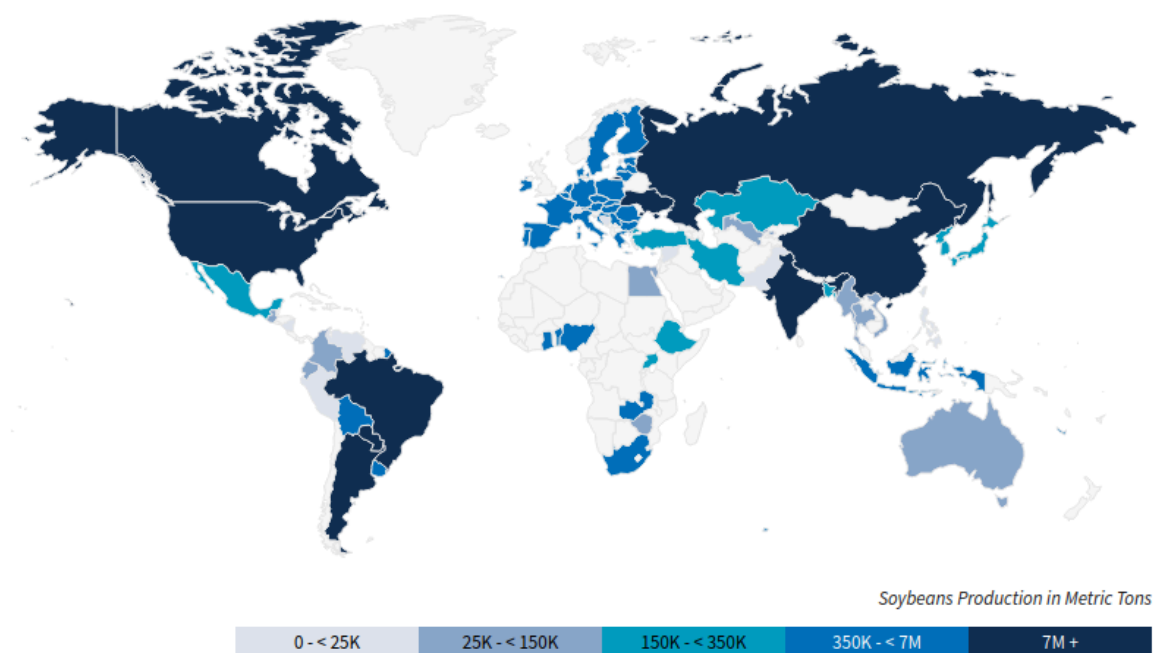


Figura 1. Distribuição global da produção de soja por país, expressa em toneladas métricas. As cores representam classes de produção, variando de menos de 25 mil toneladas até mais de 7 milhões de toneladas anuais, destacando os principais países produtores em escala global. (USDA, 2025).

Diante desse cenário, torna-se essencial compreender como os genes se diversificaram, foram conservados ou perdidos ao longo da evolução do gênero *Glycine*, identificando funções biológicas relevantes e variantes com potencial agrônomo que não estão presentes na soja cultivada. A integração entre dados de conservação gênica provenientes de espécies silvestres, perfis transcricionais derivados de RNA-Seq e informações funcionais obtidas a partir de estudos de associação genômica ampla (*Genome-wide association studies*, GWAS) oferece uma oportunidade singular para revelar genes que desempenham papéis adaptativos importantes e que podem ter sido perdidos ou divergido durante a domesticação. Portanto, este trabalho busca avançar nessa compreensão ao investigar, de forma integrada, a evolução, a conservação e a expressão de genes potencialmente estratégicos para *G. max*, avaliando como esses genes se distribuem entre linhagens cultivadas e silvestres e quais padrões de expressão podem estar associados a processos adaptativos ou agronomicamente relevantes. Com base na divergência evolutiva entre espécies anuais e perenes do gênero

Glycine, levantamos a hipótese de que genes tecido-específicos de *G. max*, especialmente aqueles associados ao desenvolvimento da semente, do embrião e de estruturas reprodutivas, apresentariam menor conservação nas espécies perenes, refletindo processos distintos de especialização funcional ao longo da evolução do gênero. Em contraste, esperava-se que genes relacionados a funções biológicas mais basais e amplamente compartilhadas, como aquelas associadas à raiz e ao nódulo, apresentassem maior conservação entre as linhagens. Assim, este trabalho buscou testar essa hipótese por meio da integração entre dados pangenômicos, perfis de expressão gênica e análises funcionais.

REVISÃO BIBLIOGRÁFICA

Diversidade genética e limitações da soja cultivada

Estima-se que a domesticação de *G. max* tenha ocorrido de 6.000 e 9.000 anos atrás, no leste da Ásia (Carter, Hymowitz e Nelson, 2004; Kim *et al.*, 2012). Esse processo envolveu a seleção de características desejáveis, como a redução da abertura espontânea das vagens durante a maturação (deiscência), sementes maiores e maior produtividade, mas também resultou em perdas substanciais de alelos raros e variantes adaptativas presentes nas populações selvagens (Hyten *et al.*, 2006).

Estudos genômicos estimam que até 50% da variabilidade genética original tenha sido perdida durante a domesticação, afetando principalmente genes associados à resistência a estresses bióticos (Bisht *et al.*, 2023), tolerância ambiental (Abdel-Ghany *et al.*, 2020) e plasticidade fenotípica (Yang *et al.*, 2024). O melhoramento genético moderno agravou ainda mais este gargalo genético (Yuan *et al.*, 2024) ao favorecer cultivares mais uniformes e com alto teor de proteína ou óleo. É consenso atualmente que ganhos futuros passam necessariamente pela incorporação, por meio de transformação ou edição gênica, da variabilidade genética encontrada em variedades semi-domesticadas e espécies selvagens do gênero *Glycine*.

Frente a esse cenário, torna-se crucial investigar o conteúdo gênico das espécies silvestres do gênero *Glycine*, que preservam ampla variabilidade genômica, visando compreender a evolução da soja e, potencialmente, para a integração de características agronômicas relevantes.

O gênero *Glycine*: diversidade, divergência e potencial evolutivo

O gênero *Glycine* é composto por dois subgêneros bem definidos; Soja, que inclui as espécies anuais *G. max* e *G. soja*; e *Glycine*, composto por 26 espécies perenes, majoritariamente endêmicas da Austrália (Newell e Hymowitz, 1983; Sherman-Broyles *et al.*, 2014). As espécies perenes apresentam grande amplitude

ecológica, ocupando ambientes variados, desde regiões áridas até florestas úmidas, o que influencia a acumulação de alelos adaptativos ausentes na soja cultivada.

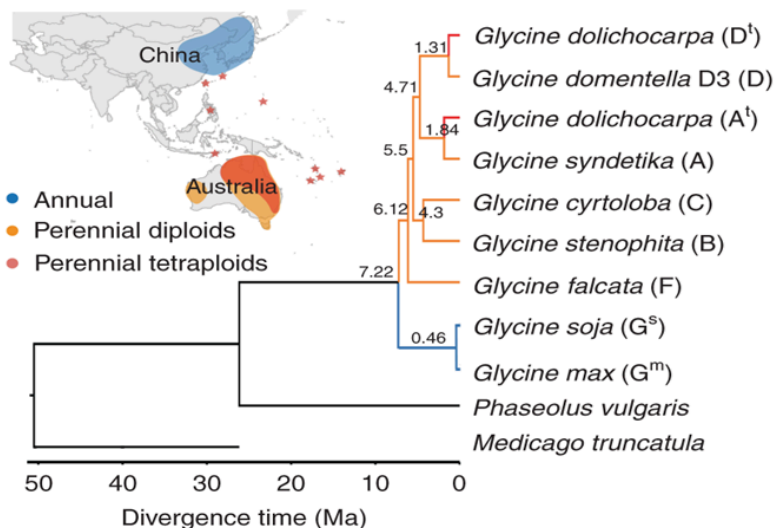


Figura 2. Relações filogenéticas, distribuição geográfica e estimativas de tempo de divergência entre espécies anuais e perenes do gênero *Glycine*. (Zhuang *et al.*, 2022)

Estima-se que os dois subgêneros tenham divergido há cerca de 5 milhões de anos, o que permitiu significativas especializações fisiológicas, estruturais e metabólicas (Kim *et al.*, 2012; Li *et al.*, 2014). Além disso, os genomas do gênero *Glycine* foram profundamente influenciados por eventos de duplicação completa do genoma (whole genome duplication, WGD), ocorridos há cerca de 59 e 13 milhões de anos (Bertioli *et al.*, 2009; Schmutz *et al.*, 2010). A WGD corresponde à duplicação integral do conteúdo cromossômico de um ancestral, aumentando o número de cópias gênicas e fornecendo matéria-prima para processos como neofuncionalização, subfuncionalização e perda diferencial de genes. Em *Glycine*, esses eventos contribuíram para a plasticidade genômica, a diversificação do gênero e a adaptação a diferentes ambientes (Koenen *et al.*, 2020; Wang *et al.*, 2023; Zhuang *et al.*, 2022). A redundância gênica resultante das WGDs não apenas proveu o material genético para a diversificação funcional, mas também forneceu uma base para a seleção de alelos favoráveis durante a domesticação e adaptação da soja cultivada.

Por outro lado, a aloploidização refere-se à formação de genomas poliploides a partir da hibridização entre linhagens ou espécies diploides distintas,

seguida da estabilização cromossômica do novo genoma híbrido. Portanto, embora tanto a WGD quanto a aloploidização possam resultar em aumento do conteúdo genético, tratam-se de processos distintos do ponto de vista evolutivo. No subgênero *Glycine*, eventos de aloploidização ocorreram nos últimos 350.000 anos, resultando na formação de aloploidos a partir da combinação de genomas diploides distintos (Bombarely, Coate e Doyle, 2014). Esse processo contribuiu para a ampliação da diversidade genética e da capacidade adaptativa das espécies, tornando o gênero *Glycine* um modelo relevante para investigar a evolução de genomas poliploides e os mecanismos de adaptação às pressões ambientais (Khan *et al.*, 2020).

As espécies perenes são particularmente interessantes por apresentarem uma maior tolerância a estresses ambientais como seca, salinidade e baixas temperaturas (Hossain, Khatoon e Komatsu, 2013), assim como uma diversidade pouco explorada de genes de resistência (genes R) que proporcionam ampla defesa contra patógenos como fungos, bactérias, vírus e nematoides (Ferreira e Marcelino-Guimarães, 2022). Compreender a diversidade genética da soja é essencial para o desenvolvimento de estratégias de melhoramento que assegurem a segurança alimentar global (Bayer *et al.*, 2022).

Além da amplificação genômica, os estágios de desenvolvimento da soja refletem uma forte integração entre a regulação genética e a resposta ambiental. Desde a germinação até a maturação das sementes, há uma sucessão coordenada de eventos fisiológicos e moleculares, mediados por redes de genes duplicados e modulados epigeneticamente (Raju *et al.*, 2018). As fases iniciais, como a formação do embrião e o desenvolvimento do cotilédono, são marcadas pela expressão de genes associados ao armazenamento de reservas e ao controle do crescimento celular (Lin *et al.*, 2017; Zhang *et al.*, 2021). Já nas fases vegetativas e reprodutivas, a duplicação gênica parece ter favorecido a especialização funcional de genes envolvidos na fotossíntese, nodulação e resistência a estresses bióticos e abióticos, que são características essenciais para a expansão ecológica e o sucesso agrônomo da espécie (Fleitas *et al.*, 2023; Lian *et al.*, 2020). Assim, compreender os estágios de desenvolvimento sob a ótica evolutiva e genômica permite

estabelecer conexões diretas com os processos de domesticação e melhoramento genético.

Nas últimas duas décadas, foi gerado um volume significativo de dados genéticos, genômicos e transcriptômicos de soja (Bayer *et al.*, 2022). Esse rápido avanço das tecnologias ômicas é crucial para o melhoramento genético de precisão, exemplificado pelo desenvolvimento de variedades de soja com composições específicas de óleo para diferentes usos (Singh *et al.*, 2008). A obtenção de genomas de referência tem sido um grande avanço, impulsionado pelo barateamento das tecnologias de sequenciamento e por novas estratégias de bioinformática. Isso abriu novas oportunidades para o melhoramento da soja por meio de ferramentas de edição gênica, como a CRISPR-Cas9 (Gaillochet, Develtere e Jacobs, 2021). Além disso, os programas de seleção assistida por marcadores (Marker-Assisted Selection) e de seleção genômica têm sido fundamentais para a identificação precisa de variantes genéticas favoráveis (Lin *et al.*, 2022). Por meio do uso de polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms*) e repetições de sequência simples (*Simple Sequence Repeats*), pode-se identificar e selecionar características genéticas específicas dentro de populações de soja (Song *et al.*, 2020). Portanto, a investigação da diversidade genética do gênero *Glycine* constitui um elemento-chave para a mineração de alelos e genes de interesse para o melhoramento da soja cultivada.

Domesticação da soja

O processo de domesticação da soja representa um marco na história da agricultura asiática e mundial. Esse processo envolveu uma série de modificações genéticas e fenotípicas que permitiram maior produtividade e adaptação ao cultivo agrícola. Características como o aumento do tamanho das sementes, a perda da deiscência das vagens e a mudança na arquitetura da planta foram fixadas por meio da seleção artificial (Sedivy, Wu e Hanzawa, 2017). Entretanto, essa seleção direcionada levou à eliminação de alelos e variantes estruturais presentes nas populações ancestrais, restringindo a base genômica da soja cultivada.

A análise da domesticação sob a perspectiva genômica tem se beneficiado do avanço de tecnologias de sequenciamento e do desenvolvimento de pangenomas, que permitem descrever o conteúdo gênico total de um conjunto de genótipos e identificar genes conservados (core) e variáveis (dispensáveis em certos contextos). Desde 2014, pangenomas de diversas espécies vegetais, incluindo arroz, milho e soja, têm revelado como a hibridização, a seleção e a adaptação moldaram a composição gênica das plantas cultivadas (Hübner *et al.*, 2019; Wang e Hu, 2021; Zhao *et al.*, 2018). No caso específico da soja, esses estudos mostraram que a domesticação alterou profundamente a proporção de genes associados à resistência a estresses bióticos e abióticos, à regulação do crescimento e à plasticidade fenotípica (Coleman *et al.*, 2021). Além disso, a comparação entre pangenomas de espécies silvestres e cultivadas do gênero *Glycine* tem demonstrado que muitas das variantes perdidas durante a domesticação permanecem preservadas nas espécies perenes, que atuam como verdadeiros reservatórios de diversidade genética. Essa diversidade é crucial para compreender a evolução e conservação funcional dos genes relacionados ao desenvolvimento, à adaptação e à tolerância ambiental. Assim, estudos da domesticação por meio de abordagens pangenômicas têm não apenas o potencial de elucidar a história evolutiva da soja, mas também de prover subsídios para o melhoramento genético sustentável, orientando o resgate de alelos de interesse presentes nas espécies silvestres.

Compreender os mecanismos e consequências genéticas da domesticação é essencial para fundamentar estratégias modernas de melhoramento genético, especialmente diante das mudanças climáticas globais e das novas condições de cultivo, que exigem a reintrodução de variabilidade genética anteriormente perdida (Zhou *et al.*, 2015). Apesar dos notáveis avanços obtidos nos programas de melhoramento, o crescimento da demanda global por alimentos, rações e biocombustíveis impõe desafios significativos. Estima-se que a produção agrícola mundial precisará dobrar até 2050 para atender ao aumento populacional, às mudanças nos hábitos alimentares e à expansão do uso de biomassa energética (Ray *et al.*, 2013). Nesse contexto, as espécies silvestres constituem uma fonte inestimável de variabilidade genética, oferecendo alelos que conferem resistência a estresses bióticos e abióticos—características frequentemente ausentes nas cultivares modernas de *G. max* (Zhao *et al.*, 2017). O aproveitamento desse

potencial genético, seja por meio de cruzamentos interespecíficos, transferência gênica ou edição genômica de precisão, é uma estratégia-chave para o desenvolvimento de variedades mais resilientes e sustentáveis frente aos desafios do futuro.

Criação do primeiro pangenoma da soja e evolução do conteúdo gênico

Tradicionalmente, os estudos genômicos se baseavam em um único genoma de referência, o que limitava o entendimento da variabilidade total de uma espécie. Com o desenvolvimento do conceito de pangenoma, tornou-se possível caracterizar o conjunto completo de genes presentes em uma população. O pangenoma inclui um genoma central contendo genes presentes em todas as cepas e um genoma acessório composto por genes ausentes em uma ou mais cepas e genes que são exclusivos de cada cepa (Tettelin *et al.*, 2005). A publicação do primeiro pangenoma de *G. max* foi realizada em 2014 (Li *et al.*, 2014), representando um marco no entendimento da diversidade genômica dessa espécie. O estudo revelou que cerca de 80% do genoma central (core genome) é conservado entre os acessos analisados, mas ainda assim evidenciou uma substancial perda de diversidade genética decorrente do processo de domesticação.

Na literatura, diferentes abordagens têm sido propostas para a construção de pangenomas (Wang *et al.*, 2023), destacando-se dois modelos principais: o pangenoma linear (Tettelin e Medini, 2020) e o pangenoma baseado em grafos (Ameur, 2019). O modelo linear é construído a partir de alinhamentos e agrupamentos filogenéticos de genes homólogos, permitindo uma descrição simplificada da presença e ausência de genes entre genótipos. Já o modelo baseado em grafos oferece uma representação mais abrangente e precisa das variações estruturais (inserções, deleções, inversões e duplicações), integrando essas diferenças em relação a um genoma de referência e possibilitando uma visão mais completa da arquitetura genômica da espécie (Wang *et al.*, 2023).

O estudo de pangenomas e super-pangenomas (i.e., pangenomas de gêneros) é, portanto, essencial para explorar a diversidade funcional do gênero *Glycine*. Em especial, ao comparar espécies anuais e perenes, torna-se possível

identificar quais genes conservados e variáveis estão associados à adaptação a diferentes ambientes, ao ciclo de vida e a características agrônômicas de interesse. Assim, o estudo do pangenoma fornece a base para análises evolutivas mais amplas, permitindo conectar a diversidade genômica à expressão gênica e à funcionalidade biológica dos genes que sustentam o sucesso adaptativo da soja e de seus parentes silvestres.

Expressão gênica e especialização tecidual

A expressão gênica corresponde ao conjunto integrado de processos moleculares que levam à utilização funcional da informação contida em um locus genômico, incluindo sua regulação, transcrição, processamento, estabilidade, tradução e a ação do produto final. Em plantas, esse processo é regulado espacial e temporalmente, de modo que muitos genes apresentam padrões transcricionais específicos em determinados tecidos, órgãos, células ou estágios do desenvolvimento. Essas especificidades são particularmente informativas, pois tendem a estar associadas a funções biológicas especializadas, como desenvolvimento embrionário, maturação de sementes, formação de tecidos de sustentação e resposta ao estresse. Todos esses processos são rigorosamente controlados por redes gênicas complexas que modulam a atividade dos genes em contextos biológicos específicos. Genes que são expressos e funcionalmente atuantes em tecidos, órgãos ou células específicos são denominados genes tecido-específicos (TS) (Liang *et al.*, 2006), e essa especificidade é geralmente inferida com base nos níveis de transcrição observados em diferentes tecidos (Sonawane *et al.*, 2017). Do ponto de vista evolutivo, investigar a conservação ou perda desses genes entre espécies anuais e perenes permite compreender quais funções são mais conservadas, quais podem ter se diversificado mais recentemente, como a especialização tecidual molda os padrões de conservação e divergência gênica e de que forma tecidos embrionários e reprodutivos acumulam divergências evolutivas (Deaner e Alper, 2018).

O presente trabalho preenche essa lacuna ao analisar a conservação e a diversificação de genes TS em múltiplas espécies do gênero *Glycine*, integrando

abordagens de pangenômica, expressão gênica e evolução funcional. Em vez de focar em genes completamente ausentes em *G. max*, este estudo concentra-se nos padrões de conservação e perda gênica entre espécies perenes e anuais, buscando compreender como a especialização tecidual influencia a retenção ou divergência de funções biológicas ao longo da evolução do gênero. A partir da identificação de genes TS conservados e não conservados, investigamos suas assinaturas funcionais por meio de análises de enriquecimento Gene Ontology, bem como a associação desses genes com processos biológicos relacionados ao desenvolvimento, metabolismo e resposta ao estresse. Adicionalmente, genes candidatos foram selecionados para revisão bibliográfica com o objetivo de contextualizar suas possíveis funções e relevância agrônoma. Essa abordagem permite revelar como a evolução da especialização tecidual contribuiu para moldar a diversidade funcional entre espécies selvagens e cultivadas de *Glycine*.

OBJETIVO GERAL

Avaliar a evolução de genes em *Glycine max* a partir de comparações de conteúdo e padrões de expressão gênica contra espécies silvestres incluindo *Glycine soja* e outras espécies do gênero

OBJETIVOS ESPECÍFICOS

- Avaliar a conservação dos genes entre a soja cultivada (*G. max*) e espécies silvestres do gênero *Glycine* silvestres, incluindo *G. soja* e espécies perenes;
- Avaliar os perfis transcricionais de genes de *G. max* conservados ou perdidos em espécies silvestres em *G. max*, visando encontrar padrões de conservação divergentes, por exemplo, entre genes tecido-específicos e housekeeping;
- Avaliar os padrões de expressão de genes com evidência de evolução adaptativa em *G. max* ou outras linhagens.

MÉTODOS

Bases de dados e conjunto genômico

Os dados utilizados neste estudo foram obtidos de duas fontes principais: o *Soybean Expression Atlas* (SEA - disponível em <https://soyatlas.venanciogroup.uenf.br/>, acesso em novembro de 2025) (Almeida-Silva, Pedrosa-Silva e Venancio, 2023), de onde obtivemos as estimativas de expressão gênica para múltiplos tecidos e uma lista de genes TS. O super-pangenoma utilizado neste trabalho foi obtido do SoyBase (disponível em <https://data.soybase.org/Glycine/GENUS/pangenomes/Glycine.pan5.MKRS/>, acesso em novembro de 2025) (Grant *et al.*, 2010).

Entre os arquivos do pangenoma, há uma matriz de contagens dos genes ortólogos em cada ortogrupo para uma seleção de genomas do gênero *Glycine*. Ao todo, são 57 genomas distribuídos em oito espécies, incluindo um genoma de cada uma das seis espécies perenes silvestres (*G. falcata*, *G. stenophita*, *G. cyrtoloba*, *G. syndetika*, *G. tomentella* e *G. dolichocarpa*), o da soja silvestre anual (*G. soja*) e 50 genomas da soja cultivada, *G. max*.

Essa matriz, disponibilizada em formato TSV pelo SoyBase, foi utilizada como base para a construção das matrizes de presença/ausência e para as análises comparativas subsequentes. Todos os dados foram processados por meio de rotinas desenvolvidas em R.

Seleção e classificação dos genes

Os genes TS foram identificados utilizando os níveis de expressão fornecidos pelo SEA. Para essa classificação, foi empregado o índice τ (tau) (Lüleci e Yilmaz, 2022), uma métrica amplamente utilizada para quantificar o grau de especificidade transcricional entre diferentes tecidos. Esse índice varia de 0 a 1, sendo que valores próximos de 0 indicam expressão ampla ou ubíqua, enquanto valores próximos de 1 indicam expressão mais restrita a um tecido específico. Neste estudo, adotou-se o valor de 0,85 como limiar mínimo de especificidade tecidual,

indicando expressão predominantemente restrita a um ou poucos tecidos. Os genes que atendiam a esse critério foram classificados de acordo com o tecido de maior expressão, resultando em uma lista final de genes TS. As informações relativas às famílias de genes ortólogos do super-pangenoma também foram obtidas do SoyBase.

Os dados foram organizados na linguagem de programação R (Giorgi, Ceraolo e Mercatelli, 2022) utilizando os pacotes *tidyverse*, *dplyr* e *tidyr* (Wickham et al., 2019). Essa abordagem metodológica possibilitou a avaliação das diferenças na conservação gênica associadas à especialização tecidual, contribuindo para a compreensão da diversificação evolutiva e funcional no gênero *Glycine*.

Esse procedimento — isto é, a integração dos genes tecido-específicos de *G. max* com as informações de presença/ausência dos respectivos ortogrupos no pangenoma — foi aplicado de forma uniforme a todas as espécies incluídas na análise, abrangendo as seis espécies perenes e em *G. soja*. Dessa forma, foi possível avaliar a conservação ou perda dos genes TS de *G. max* nas demais espécies, mantendo critérios consistentes de classificação e comparação.

Construção da matriz binária

A comparação da conservação dos genes TS de *Glycine max* entre as demais espécies do gênero *Glycine* foi realizada a partir da construção de matrizes binárias de presença/ausência baseadas nos dados do super-pangenoma. Para isso, os valores de contagem gênica por ortogrupo foram convertidos em dados binários, em que valores ≥ 1 foram codificados como “1” (presença) e valores iguais a 0 como “0” (ausência). Essa transformação permitiu avaliar se os genes TS de *G. max* estavam presentes ou ausentes nas demais espécies analisadas.

Foram construídas duas matrizes binárias independentes: uma considerando exclusivamente as seis espécies perenes silvestres e outra considerando apenas os genomas de *G. soja*. Essa separação foi adotada com base na distinção evolutiva entre as espécies perenes e a espécie anual silvestre, permitindo avaliar de forma independente os padrões de conservação e perda dos

genes TS de *G. max* em cada grupo. Dessa maneira, tornou-se possível comparar posteriormente os perfis de conservação observados nas espécies perenes com aqueles identificados em *G. soja*, considerando suas diferenças filogenéticas e seu posicionamento distinto dentro do gênero *Glycine*.

Os dados de expressão gênica utilizados na integração com a matriz binária foram exclusivamente aqueles provenientes de *G. max*, obtidos do SEA. Para cada gene de *G. max*, foi utilizada apenas a informação de classificação quanto à especificidade tecidual (TS ou não TS) e o tecido de maior expressão, não sendo utilizados os valores quantitativos de expressão. Quando um mesmo ortogruppo continha mais de um gene de *G. max*, cada gene foi considerado individualmente e associado ao seu respectivo perfil de expressão TS fornecido pelo SEA. A integração entre conservação nos ortogrupos e especificidade tecidual permitiu testar se os conjuntos de genes TS em cada tecido de *G. max* tendem a ser mais retidos ou perdidos nas espécies perenes e em *G. soja*.

Análise estatística

A associação entre a conservação gênica e a expressão tecido-específica foi avaliada por meio do Teste Exato de Fisher, a partir de tabelas de contingência 2×2 construídas com base na matriz binária de presença/ausência previamente gerada. Para essa análise, a conservação foi definida em relação à presença do ortogruppo no grupo avaliado, permitindo classificar os genes de *G. max* em categorias de conservados e não conservados conforme sua retenção nas espécies comparadas. Valores de OR > 1 indicam maior proporção de genes conservados entre os genes TS em relação aos genes não TS para um determinado tecido, enquanto valores de OR < 1 indicam maior proporção de genes não conservados entre os genes TS.

As tabelas de contingência foram então utilizadas como entrada para o Teste Exato de Fisher, que produz duas métricas principais: o valor de p, indicando a significância estatística da associação entre as variáveis, e o odds ratio (OR), que quantifica a direção e a magnitude dessa associação. Com base na hipótese do estudo, esperava-se que genes tecido-específicos associados a tecidos reprodutivos

e ao desenvolvimento da semente apresentassem maior proporção de não conservação, enquanto genes associados a funções biológicas mais basais, como aquelas relacionadas à raiz e ao nódulo, apresentassem maior proporção de conservação. Valores de OR > 1 indicam maior proporção de genes conservados entre os genes TS em relação aos genes não TS para um determinado tecido, enquanto valores de OR < 1 indicam maior proporção de genes não conservados entre os genes TS..

O teste foi aplicado separadamente para cada tecido e para cada grupo de comparação (espécies perenes e *G. soja*), permitindo avaliar se os padrões de conservação dos genes TS de *G. max* diferem entre grupos evolutivos. O limiar de significância adotado foi $p \leq 0,05$, e todas as análises foram realizadas no ambiente R.

Enriquecimento funcional

Com base nos tecidos que apresentaram associação significativa no Teste Exato de Fisher, foram selecionados nove tecidos para a realização das análises de enriquecimento funcional. O objetivo foi identificar os processos biológicos e funções moleculares mais representativos entre os genes TS de *G. max* classificados como conservados ou não conservados. A anotação funcional dos genes foi obtida a partir dados do PLAZA Dicots v5.0 (<https://bioinformatics.psb.ugent.be/plaza.dev/>) (Van Bel *et al.*, 2022). As análises de enriquecimento foram realizadas em Rstudio utilizando os pacotes do Bioconductor (Huber *et al.*, 2015), incluindo clusterProfile (Wu *et al.*, 2021), AnnotationDbi (Carlson e Gene, 2016) e do tidyverse (Wickham e Yutani, 2019). Os termos enriquecidos foram classificados em três categorias principais: Processos Biológicos, Funções Moleculares e Componentes Celulares, priorizando e selecionando apenas os processos biológicos mais relevantes. Ao final da análise, foram prospectados genes candidatos para os quais foi realizada uma revisão bibliográfica aprofundada, com o objetivo de contextualizar suas funções biológicas, possíveis implicações adaptativas e relevância para o desenvolvimento vegetal no gênero *Glycine*. Para a representação gráfica dos resultados de enriquecimento GO, os valores de significância (p-value) foram transformados para

$-\log_{10}(\text{p-value})$. Essa transformação é amplamente utilizada porque facilita a interpretação visual da significância dos termos GO enriquecidos.

RESULTADOS

A partir de dados do SEA, geramos uma lista de 3.864 genes com expressão TS, distribuídos em 17 partes da planta. Foram selecionados 9 tecidos que com maior variação no número de genes TS entre genes conservados e não conservados (Figura 2).

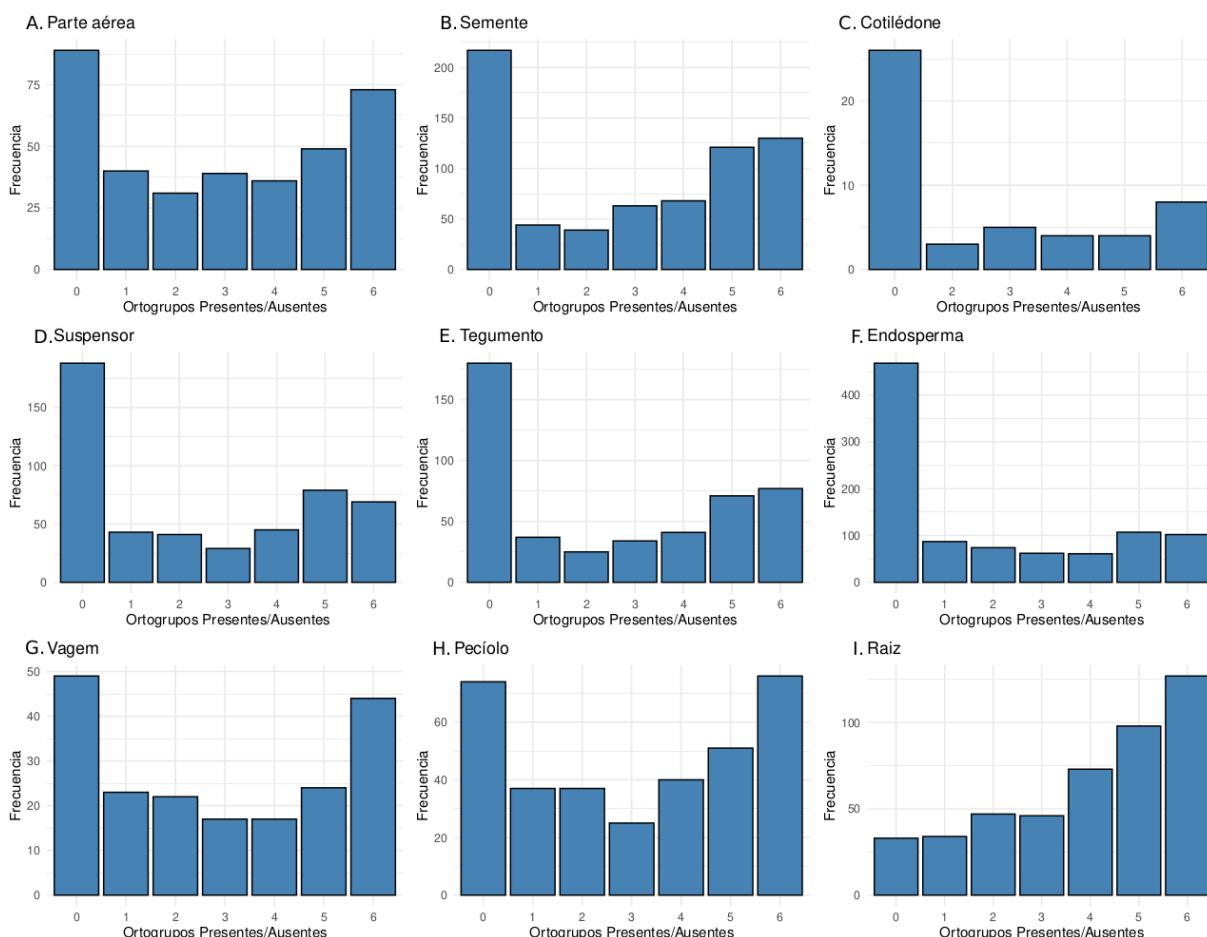


Figura 3. Padrão de conservação de genes tecido-específicos de *G. max* em espécies perenes. Os ortogrupos podem ter desde zero genes conservados em espécies perenes (exclusivos de *G. max*) à total conservação, refletida na conservação em todas as 6 espécies analisadas.

A análise do Teste Exato de Fisher (Tabela 1) permitiu corroborar essas observações, uma vez que foram avaliados 17 tecidos, dos quais oito apresentaram tendência de perda dos genes TS de *G. max* nas espécies perenes: parte aérea, semente, cotilédone, suspensor, tegumento, endosperma, vagem e pecíolo. Por outro lado, apenas um tecido, a raiz, apresentou enriquecimento de genes

conservados entre os genes TS, confirmando as tendências observadas na Figura 1. Esses resultados evidenciam que a maior parte dos tecidos analisados apresenta um padrão geral de não conservação dos genes TS de *G. max* nas espécies perenes, sugerindo que a expressão TS evoluiu após a diversificação do gênero e/ou foi selecionada artificialmente durante a domesticação e melhoramento da soja cultivada.

Essa predominância sugere que muitos genes TS expressos em estruturas como suspensor, endosperma, cotilédone e tegumento podem ter passado por processos de perda, especialização ou substituição funcional ao longo da evolução das espécies perenes. Tais tecidos estão diretamente associados ao desenvolvimento embrionário, à nutrição inicial e à formação da semente, e apresentaram elevado número de genes classificados como não conservados. (Kulmuni *et al.*, 2020; Lankau *et al.*, 2011). Em contraste, a raiz apresentou enriquecimento significativo para genes conservados, indicando que a expressão TS neste tecido evoluiu, em grande medida, antes da divergência das espécies estudadas. Considerando que a raiz desempenha funções fisiológicas essenciais — como absorção de água e nutrientes, ancoragem e percepção de estresses abióticos — é esperado que seus genes permaneçam sob maior pressão purificadora, reduzindo a taxa de perda ou divergência. Esse padrão indica que genes TS associados a tecidos vegetativos essenciais, particularmente a raiz, apresentam maior tendência à conservação entre espécies, enquanto genes TS relacionados a tecidos envolvidos no desenvolvimento e funcionalidade da semente — como suspensor, endosperma, cotilédone e tegumento — exibem maior probabilidade de ausência nas espécies perenes, refletindo maior divergência evolutiva nesses contextos funcionais. Considerando que a semente é uma das estruturas mais fortemente impactadas por processos de seleção natural e artificial, associados tanto à adaptação ambiental quanto à domesticação, é plausível que genes envolvidos nesses tecidos estejam sujeitos a maiores taxas de perda, substituição funcional ou diversificação entre linhagens do gênero *Glycine*. Esses achados estão resumidos na Tabela 1, que apresenta os valores de *p* e a direção das associações para cada tecido, permitindo identificar objetivamente quais órgãos concentram maior proporção de genes TS conservados ou não conservados de *G. max* nas demais espécies analisadas.

Tabela 1. Associação entre conservação gênica e especificidade tecidual em espécies silvestres do gênero *Glycine*.

Tecidos	TS/C	TS/NC	N-TS/C	N-TS/NC	P-valor	Odds Ratio
Parte aérea	267	85	44019	7491	$2,15 \times 10^{-06}$	0,535
Semente	464	213	43822	7363	$5,96 \times 10^{-29}$	0,366
Cotilédone	24	24	44262	7552	$7,96 \times 10^{-09}$	0,171
Suspensor	302	183	43984	7393	$1,95 \times 10^{-36}$	0,277
Tegumento	285	179	44001	7397	$3,86 \times 10^{-37}$	0,268
Endosperma	489	448	43797	7128	$1,10 \times 10^{-131}$	0,178
Vagem	145	46	44141	7530	0,001	0,538
Raiz	421	29	43865	7547	$9,11 \times 10^{-08}$	2,498
Pecíolo	265	70	44021	7506	0,002	0,646
Plântula	126	32	44160	7544	0,054	0,673
Radícula	162	39	44124	7537	0,057	0,710
Epicótilo	143	35	44143	7541	0,070	0,698
Embrião	59	17	44227	7559	0,071	0,593
Nódulo	602	115	43684	7461	0,287	0,894
Flor	493	76	43793	7500	0,438	1,111
Folha	88	17	44198	7559	0,677	0,885
Hipocótilo	38	5	44248	7571	0,828	1,300

A Tabela 1 resume os resultados obtidos pelo Teste Exato de Fisher para cada tecido analisado nas espécies silvestres de *Glycine*. Os primeiros oito tecidos correspondem aos casos em que foi detectada associação significativa ($p < 0,05$) acompanhada de odds ratio < 1 , indicando que genes tecido-específicos apresentam proporcionalmente mais casos de não conservação entre as espécies perenes. A nona célula representa um resultado com odds ratio > 2 , refletindo maior proporção de genes conservados naquele tecido em comparação aos demais. Esses padrões permitem identificar quais tecidos tendem a concentrar genes evolutivamente mais variáveis e quais apresentam maior estabilidade funcional entre as espécies analisadas.

Análise funcional e enriquecimento de genes tecido-específicos

Para caracterizar as funções biológicas associadas aos padrões de conservação identificados na etapa anterior, realizamos análises de enriquecimento funcional de termos Gene Ontology (GO) nos conjuntos de dados onde houve enriquecimento significativo. Esses genes foram selecionados porque apresentaram associação significativa no Teste Exato de Fisher e, portanto, representam os casos em que a perda gênica está potencialmente relacionada a diferenças desses tecidos entre a soja cultivada e as espécies silvestres perenes.

A análise detalhada dos termos GO mostrou que os genes TS não conservados apresentam funções particulares e coerentes com a biologia de cada tecido (Figura 3), com enriquecimento de categorias funcionais distintas entre os tecidos avaliados. Na parte aérea, observou-se enriquecimento de inibidores de peptidases e genes associados à defesa contra herbívoros e patógenos entre os genes TS não conservados, seja esperado que espécies perenes apresentem repertórios robustos de defesa, esse resultado provavelmente reflete a rápida evolução dessas famílias gênicas, caracterizadas por alta taxa de duplicação, divergência e substituição funcional. Assim, a ausência desses ortólogos específicos nas perenes não indica menor capacidade defensiva, mas sim a presença de conjuntos alternativos de genes de defesa adaptados a diferentes pressões ambientais. No cotilédone, os termos relacionados à maturação da semente e ao desenvolvimento reprodutivo refletem a alta especialização desse tecido e sua sensibilidade a pressões seletivas ambientais. Na semente, o destaque para funções de transporte ativo, glicosilação e biossíntese de ácidos graxos indica que vias metabólicas envolvidas no armazenamento energético compõem parte significativa dos genes não conservados. A vagem apresentou enriquecimento em processos mediados por etileno — hormônio-chave na maturação e deiscência — sugerindo divergência em circuitos regulatórios associados à reprodução. No pecíolo, a super-representação de atividades hidrolase relacionadas ao remodelamento da parede celular aponta para processos dinâmicos sujeitos à rápida evolução funcional. Em conjunto, esses padrões indicam que genes TS Não Conservados estão concentrados em funções associadas à especialização fisiológica e a

adaptações específicas de cada tecido, refletindo maior variabilidade evolutiva entre espécies.

Em contraste, a raiz foi o único tecido associado significativamente a genes conservados (Figura 4), revelando um padrão funcional oposto. Nesse caso, os genes TS Conservados destacaram processos essenciais, como biossíntese de fenilpropanoides, atividade peroxidase e formação da parede celular secundária — funções fundamentais para a integridade estrutural, defesa contra estresse oxidativo e manutenção da homeostase em condições ambientais diversas. A conservação dessas rotas metabólicas entre espécies perenes sugere forte pressão seletiva para preservar funções básicas e indispensáveis ao desempenho fisiológico da raiz. Assim, os resultados mostram que tecidos altamente especializados tendem a concentrar genes TS evolutivamente mais variáveis, enquanto genes TS associados à raiz — e, no caso de *G. soja*, também à flor e ao nódulo — apresentam maior tendência à conservação, refletindo a estabilidade de processos fisiológicos específicos nesses tecidos ao longo da evolução do gênero *Glycine*.

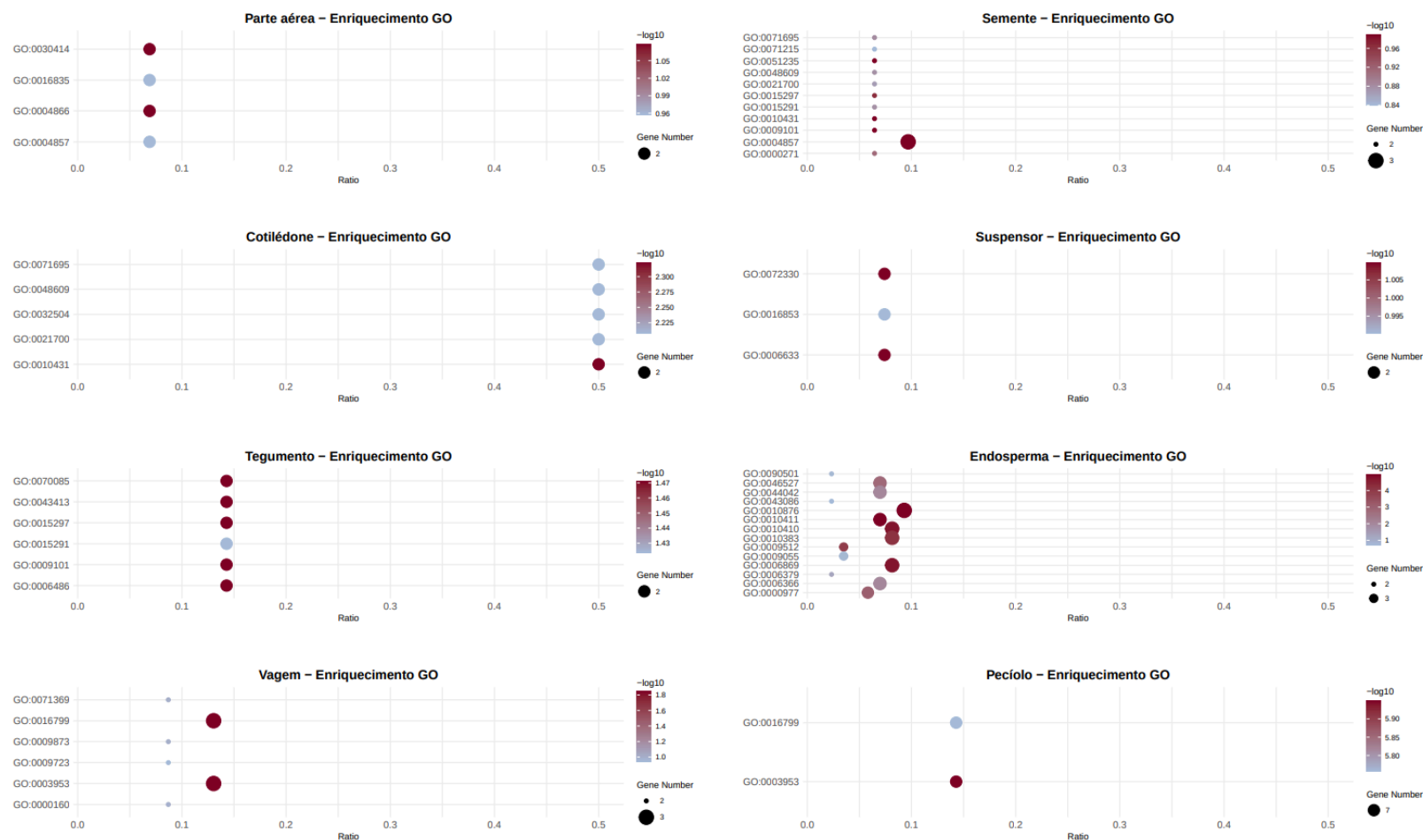


Figura 4. Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos vegetais considerando genes TS não conservados. Cada painel corresponde a um tecido. O eixo x representa a razão de enriquecimento (ratio), o tamanho dos pontos indica o número de genes associados a cada termo, e a escala de cores reflete a significância estatística do enriquecimento ($-\log_{10}$ p-valor).

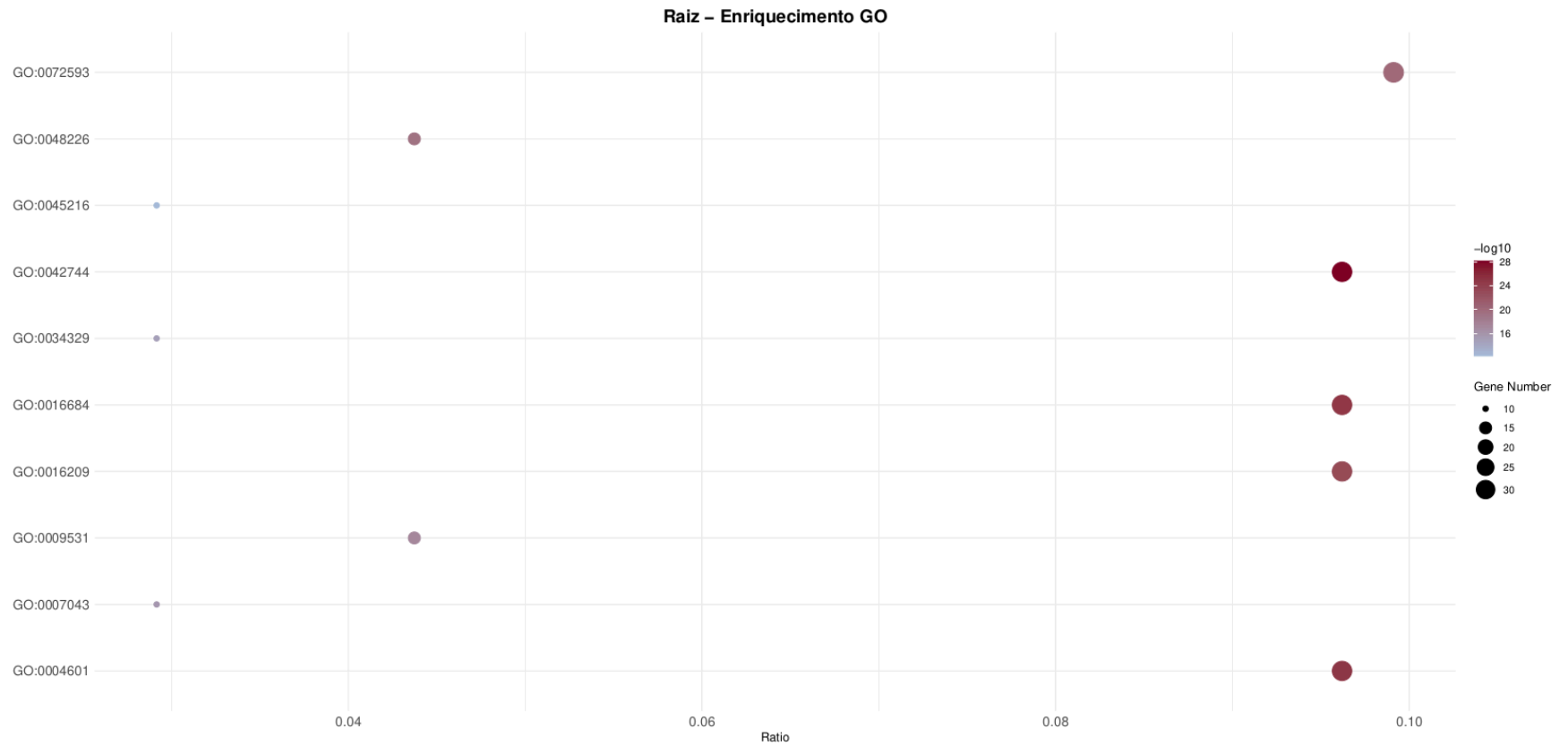


Figura 5. Análise de enriquecimento de termos Gene Ontology (GO) na raiz, considerando genes TS conservados. Cada ponto representa um termo GO enriquecido. O eixo x indica a razão de enriquecimento (ratio), o tamanho dos pontos corresponde ao número de genes associados a cada termo, e a escala de cores reflete a significância estatística do enriquecimento ($-\log_{10}$ p-valor).

No caso específico da raiz, destacaram-se processos associados ao metabolismo secundário, como a biossíntese de fenilpropanoides e a atividade peroxidase, diretamente relacionados à formação de lignina e à defesa contra estresse oxidativo. Da mesma forma, observaram-se processos de biogênese da parede celular secundária, refletindo a importância estrutural e adaptativa da raiz. Complementarmente, foi detectado um enriquecimento em genes relacionados à atividade transportadora, sugerindo um papel essencial no movimento de metabólitos e nutrientes a nível radicular. Finalmente, a presença de termos associados à resposta mediada por jasmonato e etileno indica a ativação de mecanismos de resistência sistêmica e sinalização hormonal nas raízes. Esses resultados sugerem que os genes diferencialmente expressos na raiz desempenham um papel central na adaptação estrutural, no transporte metabólico e na resposta a estresses. Com base nos termos GO significativamente enriquecidos para cada tecido, foi realizada a prospecção de genes candidatos pertencentes a essas categorias funcionais, priorizando aqueles associados a processos biológicos relevantes e com evidência de conservação nas análises comparativas. Esses genes foram então selecionados para análise individual e revisão bibliográfica, sendo apresentados na Tabela 3.

Tabela 3. Informações sobre os genes mais representativos em todos os tecidos em espécies perenes.

Tecido	Gene	Descrição	Referência (DOI)
Cotilédone	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810
Endosperma	<i>Glyma.16G162900</i>	GH16 domain-containing protein	10.1111/pbr.12357
Endosperma	<i>Glyma.20G067800</i>	GH16 domain-containing protein	10.1111/pbr.12357
Pecíolo	<i>Glyma.06G264100</i>	ADP-ribosyl cyclase/ cyclic ADP-ribose hydrolase	10.5897/JPBCS2023.1019
Pecíolo	<i>Glyma.06G265200</i>	ADP-ribosyl cyclase/ cyclic ADP-ribose hydrolase	10.5897/JPBCS2023.1019

Pecíolo	<i>Glyma.16G136000</i>	TIR domain-containing protein	10.1007/s00122-014-2438-0
Pecíolo	<i>Glyma.16G137300</i>	TIR domain-containing protein	10.1007/s00122-014-2438-0
Pecíolo	<i>Glyma.16G213800</i>	TIR domain-containing protein/ Candidate disease-resistance protein	10.1007/s00122-014-2438-0
Pecíolo	<i>Glyma.16G214700</i>	R protein	10.3389/fpls.2024.1369650
Pecíolo	<i>Glyma.16G215300</i>	NB-ARC domain-containing protein	10.3389/fpls.2024.1369650
Vagem	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810
Raiz	<i>Glyma.01G171100</i>	Peroxidase	10.1371/journal.pone.0241678
Raiz	<i>Glyma.02G201100</i>	CASP-like protein/ Casparian strip membrane protein 1	10.1007/s00299-024-03250-7
Raiz	<i>Glyma.02G231600</i>	Laccase	10.1093/g3journal/jkae226
Raiz	<i>Glyma.10G075400</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.10G075500</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.12G186500</i>	CASP-like protein/ Casparian strip membrane protein 5	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.13G315000</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.14G198900</i>	Laccase	10.1101/2020.07.25.221275
Raiz	<i>Glyma.19G083200</i>	CASP-like protein	10.1186/s12864-023-09915-z
Semente	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810
Semente	<i>Glyma.15G145600</i>	Bet v I/ Major latex protein domain-containing protein	10.1016/j.plaphy.2025.110134
Semente	<i>Glyma.16G154600</i>	Fucosyltransferase	10.1093/jxb/erw460
Tegumento	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810
Tegumento	<i>Glyma.15G145600</i>	Bet v I/ Major latex protein domain-containing protein	10.1016/j.plaphy.2025.110134
Tegumento	<i>Glyma.16G154600</i>	Fucosyltransferase	10.1093/jxb/erw460

Plântula (aérea)	<i>Glyma.01G096200</i>	Uncharacterized protein	10.3389/fpls.2023.1129454
Suspensor	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810

Na raiz, os genes candidatos identificados estão associados ao desenvolvimento estrutural e à formação de barreiras protetoras, incluindo componentes essenciais para a formação da Faixa de Caspary, uma barreira apoplástica que regula rigorosamente o fluxo de água e nutrientes do solo para a planta e restringe a entrada de substâncias potencialmente tóxicas. Entre esses genes, *Glyma.20G201000*, anotado como membro da família das caleosinas, apresentou expressão em múltiplos tecidos, sugerindo um papel funcional amplo. Estudos anteriores indicam que as caleosinas participam de processos relacionados ao desenvolvimento de sementes e à resposta a estresses abióticos (Arias *et al.*, 2022). Adicionalmente, foram identificados genes envolvidos no fortalecimento da parede celular por meio da lignificação e em mecanismos gerais de defesa, processos fundamentais para a integridade estrutural e adaptação das plantas (Zhu *et al.*, 2024).

Nos tecidos associados ao desenvolvimento da semente — semente, tegumento, endosperma e cotilédone — os genes candidatos estão majoritariamente relacionados a processos de armazenamento, desenvolvimento e metabolismo. Nesse contexto, as caleosinas desempenham papel central na formação e estabilidade dos corpos lipídicos, que constituem importantes reservas energéticas durante a germinação (Arias *et al.*, 2022). Outros genes incluem fucosiltransferases, envolvidas na modificação de carboidratos estruturais e no metabolismo de reservas como o amido, influenciando diretamente a arquitetura e funcionalidade da semente (Du *et al.*, 2017). Além disso, proteínas do tipo MLP-like têm sido associadas à resposta ao estresse em sementes (Wang *et al.*, 2025), enquanto enzimas da família GH16 atuam na modificação de β -glucanos e na remodelação da parede celular durante o desenvolvimento do endosperma e da semente, refletindo a elevada plasticidade metabólica desses tecidos (Wang *et al.*, 2016, 2021).

No pecíolo e na parte aérea, destacam-se genes do tipo R (resistência), que detectam patógenos e ativam defesas da planta, enquanto inibidores de proteases protegem fisicamente o tecido contra a digestão por pragas. Nos tecidos semente, tegumento, endosperma e cotilédono, há foco no armazenamento, desenvolvimento e metabolismo. A caleosina está presente e é importante para a formação e estabilidade dos corpos lipídicos (gorduras de reserva) (Arias *et al.*, 2022). A fucosiltransferase está envolvida na modificação de carboidratos, afetando o armazenamento de energia (amido) e a estrutura da semente (Du *et al.*, 2017). A proteína MLP-like associa-se à resistência ao estresse na semente (Wang *et al.*, 2025). Além disso, esses tecidos apresentam enzimas da família GH16 envolvidas na modificação de β -glucanos e na remodelação da parede celular durante o desenvolvimento da semente, processos essenciais para a expansão celular e organização estrutural do endosperma (Wang *et al.*, 2016, 2021).

No pecíolo e na parte aérea, destacaram-se genes do tipo R (Husein *et al.*, 2025), associados ao reconhecimento de patógenos e à ativação de respostas de defesa, bem como inibidores de proteases, que conferem proteção contra herbivoria. Na vagem, os genes identificados estão relacionados a processos de maturação e resposta ao estresse, incluindo a atuação de caleosinas em contextos associados à dessecação e à defesa pré-germinativa, possivelmente modulados por vias de sinalização hormonal como o etileno. De forma integrada, os resultados da análise de enriquecimento GO e da prospecção direcionada de genes demonstram concordância funcional, permitindo a identificação de genes candidatos relevantes para estudos futuros. Esses genes apresentam potencial para aplicações em melhoramento genético, especialmente no aumento da resistência a estresses, no aprimoramento do desenvolvimento de tecidos-chave como raiz e semente, e na caracterização funcional de genes ainda pouco explorados, destacando a importância das caleosinas na fisiologia vegetal.

Após as análises conduzidas para as espécies perenes, foi realizada uma segunda etapa de enriquecimento funcional considerando exclusivamente *G. soja*, com o objetivo de avaliar se os padrões observados nas espécies ancestrais se mantêm na linhagem anual silvestre. Para isso, foram analisados separadamente os

genes tecido-específicos classificados como Não Conservados (Figura 4) e Conservados (Figura 5).

De forma complementar à análise realizada com as espécies perenes, o Teste Exato de Fisher foi aplicado aos genes TS de *G. max* em relação à conservação em *G. soja*. Essa análise revelou um padrão distinto, no qual raiz, flor e nódulo apresentaram associação significativa com genes conservados, enquanto os demais tecidos mostraram tendência à não conservação (Tabela 2). Com base nesses resultados, os tecidos com associação significativa foram selecionados para as análises de enriquecimento funcional descritas a seguir.

Tecidos	TS/C	TS/NC	N-TS/C	N-TS/NC	P-valor	Odds Ratio
Semente	552	125	47229	3956	$3,24 \times 10^{-19}$	0,37
Cotilédone	35	13	47746	4068	$6,08 \times 10^{-5}$	0,229
Suspensor	350	135	47431	3946	$1,67 \times 10^{-5}$	0,216
Tegumento	355	109	47426	3972	$2,22 \times 10^{-25}$	0,273
Endosperma	671	266	47110	3815	$3,69 \times 10^{-79}$	0,204
Radícula	175	26	47606	4055	0,012	0,573
Parte aérea	309	43	47472	4038	0,004	0,611
Raiz	443	7	47338	4074	$5,59 \times 10^{-9}$	5,446
Flor	546	23	47235	4058	0,001	2,039
Nódulo	679	38	47102	4043	0,01	1,534
Folha	102	3	47679	4078	0,066	2,908
Plântula	140	18	47641	4063	0,103	0,663
Embrião	67	9	47714	4072	0,198	0,635
Epicótilo	168	10	47613	4071	0,328	1,436
Pecíolo	312	23	47469	4058	0,61	1,16
Hipocótilo	41	2	47740	4079	0,773	1,751
Vagem	176	15	47605	4066	1	1,002

Tabela 2. Associação entre conservação gênica e especificidade tecidual em *G. soja*.

Apresentam-se os valores obtidos pelo Teste Exato de Fisher para diferentes tecidos da *G. soja*.

Enriquecimento GO dos genes TS em *G. max* e perdidos em *Glycine soja*

A análise de enriquecimento funcional dos genes TS Não Conservados em *G. soja* revelou padrões funcionais distintos entre os diferentes tecidos, indicando que a perda ou divergência gênica ocorre de forma específica para cada órgão.

Os genes não conservados da semente estão associados a processos de transporte ativo, atividade enzimática regulatória e metabolismo energético, refletindo funções essenciais para acúmulo de reservas e transição para germinação. Esses processos parecem ter se diversificado ao longo da evolução, sugerindo adaptação a condições ambientais variadas. O cotilédone apresentou termos GO relacionados ao desenvolvimento embrionário, metabolismo de lipídios e processamento de proteínas, indicando que funções ligadas à nutrição inicial e formação do embrião exibem maior plasticidade evolutiva. No suspensor, observaram-se funções associadas ao transporte de solutos, regulação redox e processos metabólicos iniciais, coerentes com seu papel transiente no início do desenvolvimento embrionário. A divergência desses genes sugere pressões seletivas específicas que atuam nas fases iniciais do desenvolvimento. O tegumento mostrou enriquecimento em termos relacionados à defesa química, modificação da parede celular e atividade hidrolítica, indicando que processos de proteção da semente são altamente sujeitos à diversificação, possivelmente devido a adaptações a diferentes predadores e ambientes. No endosperma, os termos enriquecidos envolveram metabolismo energético, resposta hormonal e produção de reservas, reforçando que esse tecido acumula variação funcional ligada à nutrição embrionária. Os genes não conservados da parte aérea destacaram processos associados à resposta hormonal, regulação do crescimento e defesa contra estresses bióticos, indicando que funções acima do solo são mais sensíveis à variação ambiental e evolutiva. A radícula apresentou enriquecimento em termos relacionados à atividade oxidativa, hidrólise de compostos orgânicos e remodelação da parede celular, sugerindo que funções iniciais de emergência da raiz também passaram por modificações evolutivas relevantes.

Em conjunto, os genes TS de *G. max* não conservados em *G. soja* estão principalmente associados a processos relacionados ao desenvolvimento da

semente, metabolismo energético, remodelação da parede celular e resposta hormonal. Considerando a proximidade evolutiva entre *G. soja* e *G. max*, a ausência desses genes indica divergência funcional recente, particularmente em tecidos embrionários, em contraste com o padrão observado nas espécies perenes, nas quais apenas a raiz apresentou associação consistente com conservação gênica. Além disso, em *G. soja* observou-se conservação adicional em flor e nódulo, sugerindo que processos ligados à reprodução e à simbiose foram mantidos após a separação dessas linhagens. Esses resultados indicam que genes TS associados a tecidos da semente exibem elevada rotatividade entre *G. max* e *G. soja*.

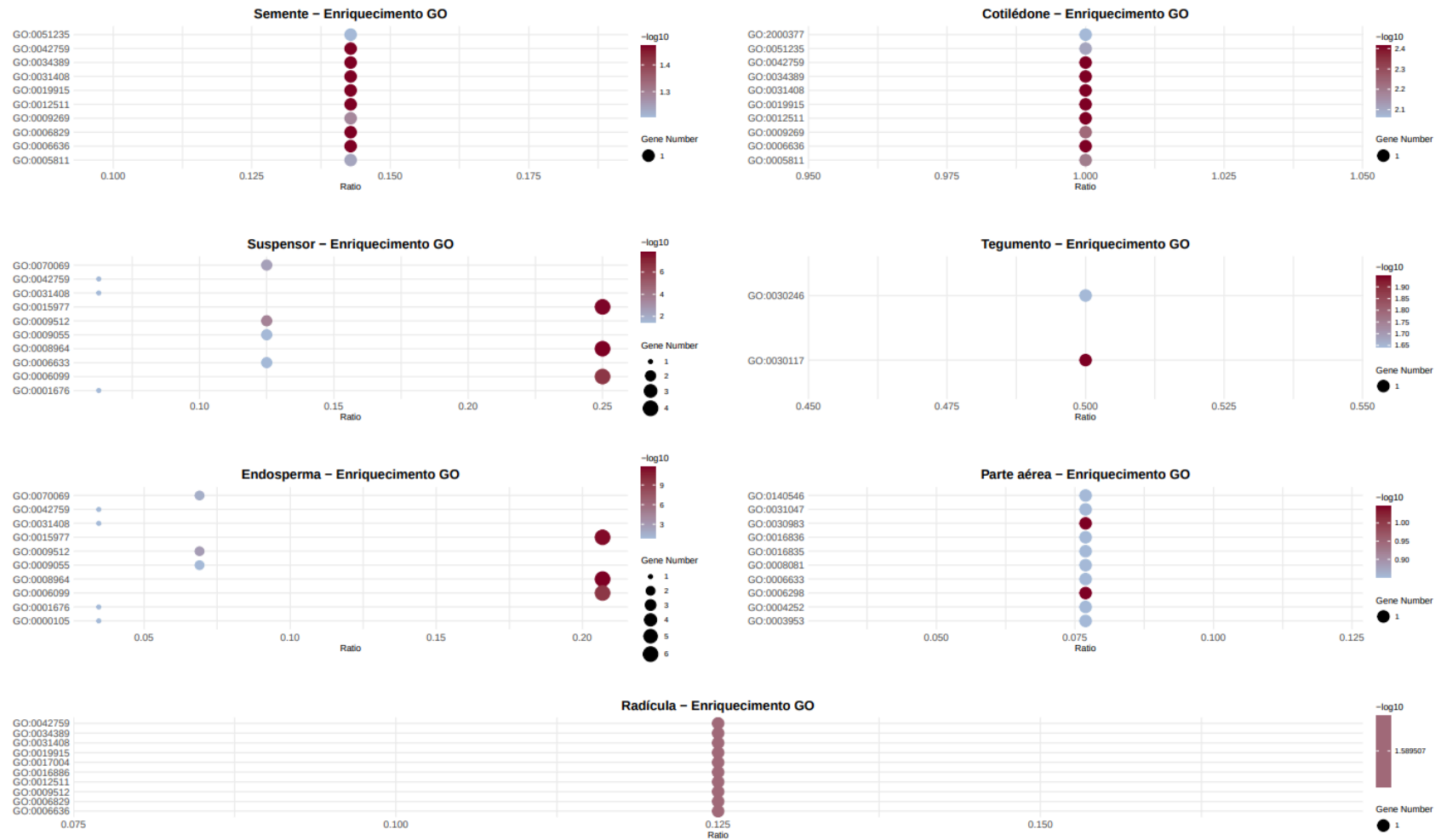


Figura 6. Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos de Glycine soja, considerando genes não conservados. Cada painel corresponde a um tecido específico. O eixo x representa a razão de enriquecimento (ratio), o tamanho dos pontos indica o número de genes associados a cada termo GO, e a escala de cores reflete a significância estatística do enriquecimento ($-\log_{10} p$ -valor).

Enriquecimento GO dos genes TS em *G. max* e conservados em *Glycine soja*

A análise dos genes TS em *G. max* e conservados em *G. soja* revelou apenas três tecidos com enriquecimento significativo; Raiz, flor e nódulo. Esses tecidos exibem funções mais estáveis ao longo da evolução. A raiz apresentou enriquecimento em vias relacionadas à biossíntese de fenilpropanoides, atividade peroxidase, formação da parede celular secundária e respostas oxidativas. Esses processos são essenciais para sustentação, absorção e defesa contra estresses do solo, justificando seu alto grau de conservação. A flor apresentou termos associados à regulação hormonal, sinalização celular e metabolismo energético, indicando que funções ligadas à reprodução mantêm estabilidade evolutiva, possivelmente devido a restrições funcionais impostas pelo sucesso reprodutivo. Os genes conservados do nódulo estão vinculados a processos relacionados à fixação biológica de nitrogênio, metabolismo redox e respostas a estresses, reforçando a importância adaptativa desse tecido em leguminosas e a necessidade de manter funções simbióticas essenciais. Ademais, nossos resultados sugerem que a expressão TS de tais genes evoluiu em grande medida anteriormente à domesticação da soja cultivada.

Os genes TS conservados em *G. soja* representam funções essenciais e estruturalmente críticas, especialmente na raiz e no nódulo, enquanto a flor mantém estabilidade associada à reprodução. Esses padrões contrastam com a maior variabilidade observada nos tecidos analisados anteriormente.

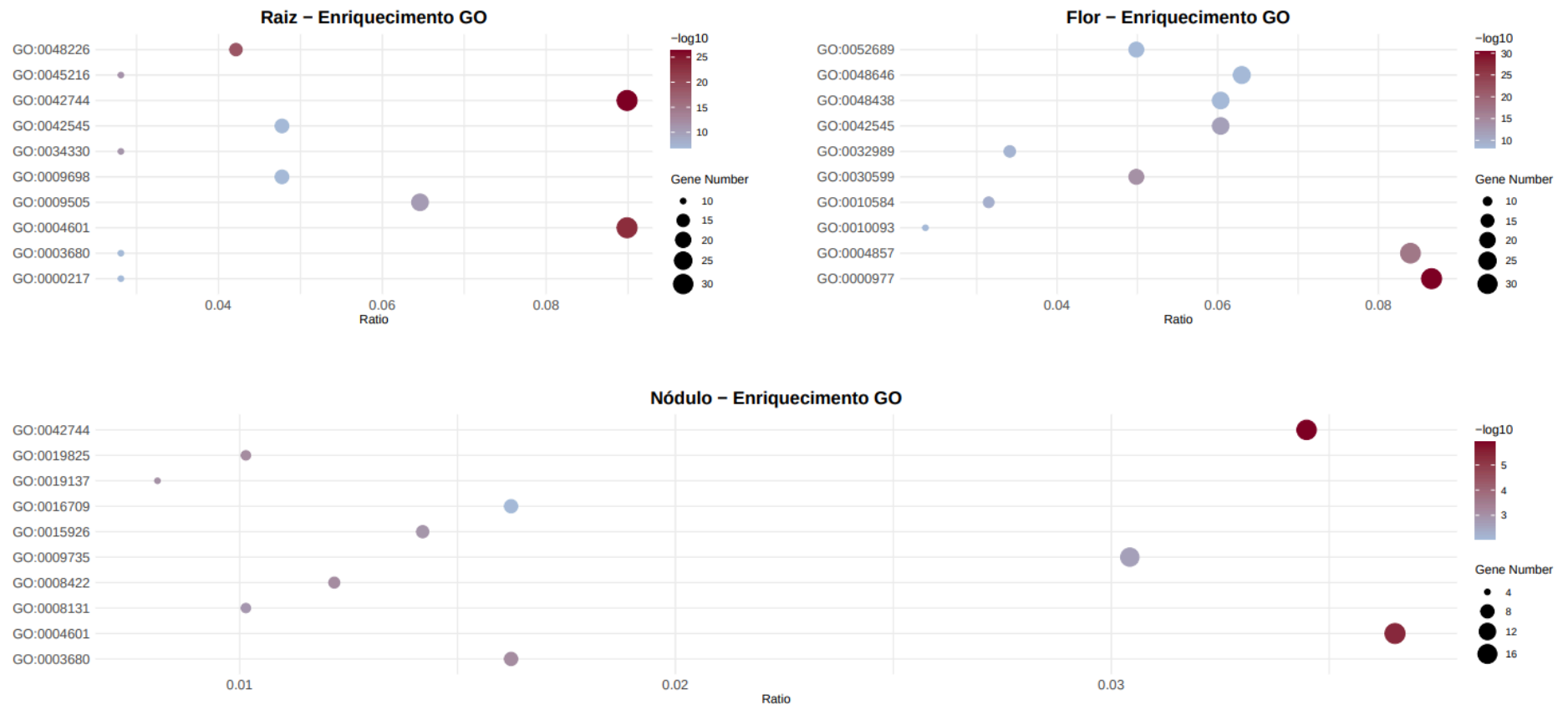


Figura 7. Análise de enriquecimento de termos Gene Ontology (GO) em diferentes tecidos de *Glycine soja*, considerando genes conservados. Cada painel corresponde a um tecido específico (raiz, flor e nódulo). O eixo x representa a razão de enriquecimento (ratio), o tamanho dos pontos indica o número de genes associados a cada termo GO, e a escala de cores reflete a significância estatística do enriquecimento ($-\log_{10}$ p-valor).

Com base no enriquecimento significativo de termos GO na análise envolvendo *G. soja*, foi realizada a prospecção de genes candidatos pertencentes a categorias funcionais associadas ao desenvolvimento, metabolismo e resposta ao estresse. A seleção priorizou genes pertencentes a termos GO significativamente enriquecidos e biologicamente relevantes para os tecidos analisados. Esses genes foram então submetidos a uma revisão bibliográfica direcionada, com o objetivo de reunir evidências funcionais que sustentem seu potencial para futuros estudos de caracterização molecular e aplicações em programas de melhoramento genético, especialmente como fontes de diversidade funcional presentes em *G. soja*. A revisão da literatura indica que esses genes participam de processos relevantes para adaptação e desempenho agrônomo. As informações consolidadas para esses genes são apresentadas na Tabela 4, incluindo a função biológica descrita e as respectivas referências bibliográficas.

Tabela 4. Informações sobre os genes mais representativos em todos os tecidos de *G. soja*.

Tecido	Gene	Descrição	Referência
Semente/Cotilédone/Suspensor/Endosperma/Radícula	<i>Glyma.20G201000</i>	EF-hand domain-containing protein	10.1111/pbi.13810
Suspensor/Endosperma/Radícula	<i>Glyma.09G171300</i>	Codifica a subunidade 8 do complexo citocromo b6-f.	10.3389/fpls.2023.1229495
Raiz/Nódulo	<i>Glyma.01G171100</i>	Peroxidase superfamily protein	10.1371/journal.pone.0241678
Raiz/Nódulo	<i>Glyma.12G195600</i>	Peroxidase 39-like	10.1093/jxb/erx441
Raiz	<i>Glyma.02G201100</i>	CASP-like protein/Casparian strip membrane protein 1	10.1007/s00299-024-03250-7
Raiz	<i>Glyma.10G075400</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.10G075500</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.12G186500</i>	CASP-like protein/ Casparian strip membrane protein 5	10.1186/s12864-023-09915-z
Raiz	<i>Glyma.13G315000</i>	CASP-like protein	10.1186/s12864-023-09915-z
Raiz/Nódulo	<i>Glyma.17G052900</i>	Peroxidase 59	10.1101/2025.07.03.662952
Raiz/Nódulo	<i>Glyma.17G053000</i>	Peroxidase 59	10.1101/2025.07.03.662952

DISCUSSÃO

A evolução do gênero *Glycine* é marcada por diferenças expressivas na probabilidade de conservação de genes TS de *G. max* entre espécies silvestres perenes e a espécie anual *G. soja*. Os resultados deste estudo indicam que, para a maioria dos tecidos analisados, os genes TS de *G. max* apresentam maior tendência à perda nas espécies perenes do que esperado ao acaso, enquanto a raiz se destaca como o único tecido cujos genes TS tendem a ser conservados. Esse padrão sugere que os padrões de expressão TS em *G. max* surgiram após a divergência do gênero e foram importantes para a retenção de tais genes.

Nos tecidos associados ao desenvolvimento e à proteção da semente — como cotilédone, suspensor, endosperma, tegumento e vagem — observou-se que os genes TS de *G. max* apresentam maior probabilidade de estarem ausentes nas espécies perenes. Esses genes mostraram enriquecimento em processos relacionados ao desenvolvimento embrionário, armazenamento de reservas, metabolismo energético, regulação hormonal e defesa química (Du *et al.*, 2017; Wang *et al.*, 2025). Tais processos estão diretamente envolvidos na formação e funcionalidade da semente, uma estrutura fortemente sujeita tanto à seleção natural quanto à seleção artificial durante a domesticação (Meyer e Purugganan, 2013). Esse contexto sugere que genes associados a esses tecidos podem ter sido diferencialmente retidos, substituídos ou perdidos em diferentes linhagens do gênero *Glycine*, refletindo adaptações a ambientes distintos e estratégias reprodutivas contrastantes entre espécies perenes e anuais.

Em contraste, a raiz foi o único tecido em que os genes TS de *G. max* apresentaram maior tendência de conservação nas espécies perenes. O enriquecimento funcional observado nesse tecido incluiu processos associados à lignificação, formação de parede celular secundária, metabolismo oxidativo e biossíntese de fenilpropanoides, funções essenciais para a integridade estrutural, absorção de recursos e resposta a estresses abióticos. Esses resultados sugerem que genes TS associados a funções fisiológicas fundamentais tendem a estar sob maior pressão purificadora (Chairom *et al.*, 2024; Mansoor *et al.*, 2024; Zhu *et al.*, 2024), o que pode reduzir a probabilidade de perda gênica ao longo da evolução. No

entanto, esse padrão foi particularmente evidente na raiz, tecido em que foram observados enriquecimentos relacionados a processos estruturais, metabólicos e de resposta a estresses. Assim, embora a expressão seja tecido-específica, a conservação desses genes pode refletir a importância funcional dessas rotas metabólicas para a manutenção do desempenho fisiológico das plantas em diferentes condições ambientais.

A análise independente envolvendo *G. soja* revelou um padrão parcialmente distinto daquele observado nas espécies perenes. Além da raiz, os tecidos de flor e nódulo apresentaram maior proporção de genes TS de *G. max* classificados como conservados em *G. soja*. Esses tecidos estão associados a processos biologicamente críticos, como reprodução e simbiose com rizóbios, respectivamente. A conservação de genes TS nesses contextos sugere que pressões seletivas relacionadas à fertilidade e à fixação biológica de nitrogênio podem favorecer a manutenção de genes específicos nesses tecidos no clado das espécies anuais (Mansoor *et al.*, 2024; Wu *et al.*, 2018). A ausência desses padrões na análise das espécies perenes indica que diferentes trajetórias evolutivas e estratégias de vida podem moldar de forma distinta a evolução da expressão TS e a retenção de tais genes entre grupos do gênero *Glycine*.

Por outro lado, para os demais tecidos em *G. soja*, os genes TS de *G. max* também apresentaram maior probabilidade de estarem ausentes, com enriquecimento funcional em processos metabólicos, transporte ativo, biossíntese de ácidos graxos e rotas de defesa e sinalização. Esses resultados indicam que tecidos funcionalmente especializados e metabolicamente ativos continuam sendo contextos associados a maior variabilidade gênica, mesmo após a domesticação. Assim, embora a domesticação tenha imposto gargalos genéticos, parte significativa da variabilidade funcional relacionada a esses tecidos permanece diferenciada entre *G. max* e *G. soja*.

Em conjunto, os resultados indicam que a evolução funcional no gênero *Glycine* reflete um equilíbrio entre estabilidade e diversificação: genes TS associados a funções fisiológicas essenciais tendem a apresentar maior probabilidade de conservação entre espécies, enquanto genes TS relacionados a

tecidos envolvidos no desenvolvimento da semente, proteção do embrião e metabolismo especializado apresentam maior probabilidade de ausência em espécies silvestres, refletindo maior divergência funcional entre linhagens (Zhou *et al.*, 2015). Esse padrão reforça a importância de considerar a especificidade tecidual ao investigar processos de conservação gênica em análises pangenômicas comparativas.

Outro aspecto relevante deste estudo é o papel das espécies perenes como importantes reservatórios de variantes funcionais. A elevada proporção de genes TS de *G. max* ausentes nas espécies perenes indica divergência funcional significativa entre linhagens, sugerindo que parte desses genes pode ter sido adquirida ou fixada ao longo do processo de domesticação e melhoramento da soja cultivada, refletindo adaptações específicas ao ambiente agrícola. Essa diversidade representa um recurso valioso para programas de melhoramento genético, especialmente para a introdução de características associadas à tolerância a estresses, adaptação ambiental e plasticidade metabólica.

Finalmente, ao integrar dados de pangenoma, expressão tecido-específica e análises funcionais, este estudo oferece uma visão abrangente da evolução funcional no gênero *Glycine*. Os padrões de conservação e ausência de genes TS de *G. max* observados entre espécies silvestres revelam como pressões seletivas específicas associadas a diferentes contextos funcionais moldaram a trajetória evolutiva do grupo, destacando conjuntos gênicos potencialmente associados à adaptação da soja cultivada ao ambiente agrícola. Esses genes constituem alvos relevantes para estudos funcionais e estratégias de edição genômica e melhoramento assistido por marcadores, visando otimizar características agronômicas específicas da soja moderna (Sim *et al.*, 2023).

CONCLUSÕES

- Este estudo integrou dados de pangenoma, expressão TS e análises de enriquecimento funcional para avaliar a probabilidade de conservação ou perda de genes TS de *G. max* em espécies silvestres perenes e em *G. soja*.
- Os resultados demonstraram que, nas espécies perenes, a maioria dos tecidos apresenta maior probabilidade de ausência de genes TS de *G. max*, especialmente aqueles associados ao desenvolvimento embrionário, formação e proteção da semente, metabolismo energético, defesa química e remodelação estrutural.
- A raiz destacou-se como o único tecido em que os genes TS de *G. max* tendem a ser mais frequentemente conservados nas espécies perenes, refletindo a importância evolutiva de funções estruturais e fisiológicas essenciais, como lignificação, metabolismo oxidativo e biossíntese de fenilpropanoides.
- A análise independente em *G. soja* revelou um padrão parcialmente distinto, no qual, além da raiz, os tecidos de flor e nódulo apresentaram maior proporção de genes TS conservados, sugerindo que os padrões TS em tais tecidos precederam a domesticação da soja cultivada.
- Em contraste, tecidos metabolicamente especializados em *G. soja*, como semente, cotilédono e radícula, apresentaram maior probabilidade de ausência de genes TS de *G. max*, indicando que processos ligados ao desenvolvimento da semente e ao metabolismo especializado continuam sendo contextos associados à maior divergência funcional.
- As espécies perenes evidenciam trajetórias evolutivas distintas em relação à soja cultivada, refletindo divergência funcional significativa entre linhagens e sugerindo a fixação de genes TS específicos ao longo da domesticação da soja.
- A identificação de famílias gênicas como caleosinas, peroxidases, laccases, proteínas CASP-like e membros da família TIR-NB-ARC destaca conjuntos de genes candidatos que poderão ser explorados em estudos funcionais futuros

e em estratégias de melhoramento assistido por marcadores ou edição genômica.

- Este trabalho mostra que abordagens integrativas baseadas em pangenoma, expressão TS e análises evolutivas são fundamentais para compreender a evolução do gênero *Glycine* e fornecendo subsídios para o desenvolvimento de novas cultivares de soja mais resilientes e com maior base genética.

REFERÊNCIAS BIBLIOGRÁFICAS

- ABDELGHANY, A. M. *et al.* Natural Variation in Fatty Acid Composition of Diverse World Soybean Germplasms Grown in China. **Agronomy**, v. 10, n. 1, p. 24, jan. 2020.
- ABDEL-GHANY, S. E. *et al.* Transcriptome Analysis of Drought-Resistant and Drought-Sensitive Sorghum (*Sorghum bicolor*) Genotypes in Response to PEG-Induced Drought Stress. **International Journal of Molecular Sciences**, v. 21, n. 3, 23 jan. 2020.
- ALMEIDA-SILVA, F.; PEDROSA-SILVA, F.; VENANCIO, T. M. The Soybean Expression Atlas v2: A comprehensive database of over 5000 RNA-seq samples. **The Plant Journal: For Cell and Molecular Biology**, 8 set. 2023.
- AMEUR, A. Goodbye reference, hello genome graphs. **Nature Biotechnology**, v. 37, n. 8, p. 866–868, ago. 2019.
- ARIAS, C. L. *et al.* Expression of AtWRI1 and AtDGAT1 during soybean embryo development influences oil and carbohydrate metabolism. **Plant Biotechnology Journal**, v. 20, n. 7, p. 1327–1345, 2022.
- BAYER, P. E. *et al.* Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. **The Plant Genome**, v. 15, n. 1, p. e20109, 2022.
- BERTIOLI, D. J. *et al.* An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. **BMC Genomics**, v. 10, n. 1, p. 45, 23 jan. 2009.
- BISHT, A. *et al.* Multi-omics assisted breeding for biotic stress resistance in soybean. **Molecular Biology Reports**, v. 50, n. 4, p. 3787–3814, abr. 2023.
- BOMBARELY, A.; COATE, J. E.; DOYLE, J. J. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. **PeerJ**, v. 2, p. e391, 20 maio 2014.
- CARLSON, M.; GENE, I. D. **AnnotationDbi: Introduction To Bioconductor Annotation Packages** 2016 Disponível em: <<https://www.semanticscholar.org/paper/AnnotationDbi%3A-Introduction-To-Bioconductor-Carlson-Gene/7daaab84c7be3e3e071b33d58bbc15caf6e90944>>. Acesso em: 9 jan. 2026
- CARTER, T. E.; HYMOWITZ, T.; NELSON, R. L. **Biogeography, Local Adaptation, Vavilov, and Genetic Diversity in Soybean** (D. Werner, Ed.) Biological Resources and Migration. **Anais...** Berlin, Heidelberg: Springer, 2004
- CHAIPROM, U. *et al.* Impact of Rhg1 copy number variation on a soybean cyst nematode resistance transcriptional network. **G3 Genes|Genomes|Genetics**, v. 14, n. 12, p. jkae226, 1 dez. 2024.

COLEMAN, K. *et al.* The potential for soybean to diversify the production of plant-based protein in the UK. **The Science of the Total Environment**, v. 767, p. 144903, 1 maio 2021.

DEANER, M.; ALPER, H. S. Promoter and Terminator Discovery and Engineering. **Advances in Biochemical Engineering/Biotechnology**, v. 162, p. 21–44, 2018.

DU, J. *et al.* Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. **Journal of Experimental Botany**, v. 68, n. 8, p. 1955–1972, 1 abr. 2017.

FERREIRA, E. G. C.; MARCELINO-GUIMARÃES, F. C. Mapping Major Disease Resistance Genes in Soybean by Genome-Wide Association Studies. **Methods in Molecular Biology (Clifton, N.J.)**, v. 2481, p. 313–340, 2022.

FLEITAS, A. L. *et al.* Functional specialization of chloroplast vesiculation (CV) duplicated genes from soybean shows partial overlapping roles during stress-induced or natural senescence. **Frontiers in Plant Science**, v. 14, p. 1184020, 6 jun. 2023.

GAILLOCHET, C.; DEVELTERE, W.; JACOBS, T. B. CRISPR screens in plants: approaches, guidelines, and future prospects. **The Plant Cell**, v. 33, n. 4, p. 794–813, 5 abr. 2021.

GIORGI, F. M.; CERAULO, C.; MERCATELLI, D. The R Language: An Engine for Bioinformatics and Data Science. **Life (Basel, Switzerland)**, v. 12, n. 5, p. 648, 27 abr. 2022.

GRAHAM, P. H.; VANCE, C. P. Legumes: importance and constraints to greater use. **Plant Physiology**, v. 131, n. 3, p. 872–877, mar. 2003.

GRANT, D. *et al.* SoyBase, the USDA-ARS soybean genetics and genomics database. **Nucleic Acids Research**, v. 38, n. Database issue, p. D843-846, jan. 2010.

HOSSAIN, Z.; KHATOON, A.; KOMATSU, S. Soybean proteomics for unraveling abiotic stress response mechanism. **Journal of Proteome Research**, v. 12, n. 11, p. 4670–4684, 1 nov. 2013.

HUBER, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, v. 12, n. 2, p. 115–121, fev. 2015.

HÜBNER, S. *et al.* Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. **Nature Plants**, v. 5, n. 1, p. 54–62, jan. 2019.

HUSEIN, G. *et al.* **Transcriptome Profiling of Resistance Genes Analogs in Soybean's Cross-Tolerance to Water Limitation and Rust Stress** bioRxiv, , 23 maio 2025. Disponível em: <<https://www.biorxiv.org/content/10.1101/2025.05.20.652543v1>>. Acesso em: 2 fev. 2026

HYTEN, D. L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 45, p. 16666–16671, 7 nov. 2006.

KHAN, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. **Trends in Plant Science**, v. 25, n. 2, p. 148–158, fev. 2020.

KIM, M. Y. *et al.* Tracing soybean domestication history: From nucleotide to genome. **Breeding Science**, v. 61, n. 5, p. 445–452, jan. 2012.

KOENEN, E. J. M. *et al.* Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. **New Phytologist**, v. 225, n. 3, p. 1355–1369, 2020.

KULMUNI, J. *et al.* Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 375, n. 1806, p. 20190528, 13 jul. 2020.

LANKAU, R. *et al.* Incorporating evolutionary principles into environmental management and policy. **Evolutionary Applications**, v. 4, n. 2, p. 315–325, mar. 2011.

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

LI, YING-HUI *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. **Nature Biotechnology**, v. 32, n. 10, p. 1045–1052, out. 2014.

LIAN, S. *et al.* The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences. **BMC Plant Biology**, v. 20, p. 277, 16 jun. 2020.

LIANG, S. *et al.* Detecting and profiling tissue-selective genes. **Physiological Genomics**, v. 26, n. 2, p. 158–162, 12 jul. 2006.

LIN, F. *et al.* Breeding for disease resistance in soybean: a global perspective. **TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik**, v. 135, n. 11, p. 3773, 5 jul. 2022.

LIN, J.-Y. *et al.* Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 45, p. E9730–E9739, 7 nov. 2017.

LÜLECI, H. B.; YILMAZ, A. Robust and rigorous identification of tissue-specific genes by statistically extending tau score. **BioData Mining**, v. 15, p. 31, 9 dez. 2022.

MANSOOR, S. *et al.* Comparative transcriptomic analysis of the nodulation-competent zone and inference of transcription regulatory network in silicon applied *Glycine max* [L.]-Merr. Roots. **Plant Cell Reports**, v. 43, n. 7, p. 169, 12 jun. 2024.

MEYER, R. S.; PURUGGANAN, M. D. Evolution of crop species: genetics of domestication and diversification. **Nature Reviews Genetics**, v. 14, n. 12, p. 840–852, dez. 2013.

NEWELL, C. A.; HYMOWITZ, T. Hybridization in the Genus Glycine Subgenus Glycine Willd. (leguminosae, Papilionoideae). **American Journal of Botany**, v. 70, n. 3, p. 334–348, 1983.

RAJU, S. K. K. *et al.* An epigenetic breeding system in soybean for increased yield and stability. **Plant Biotechnology Journal**, v. 16, n. 11, p. 1836–1847, 2018.

RAY, D. K. *et al.* Yield Trends Are Insufficient to Double Global Crop Production by 2050. **PloS One**, v. 8, n. 6, p. e66428, 2013.

RAZZAQ, A. *et al.* Next-Generation Breeding Strategies for Climate-Ready Crops. **Frontiers in Plant Science**, v. 12, 21 jul. 2021.

SCHMUTZ, J. *et al.* Genome sequence of the palaeopolyploid soybean. **Nature**, v. 463, n. 7278, p. 178–183, jan. 2010.

SEDIVY, E. J.; WU, F.; HANZAWA, Y. Soybean domestication: the origin, genetic architecture and molecular bases. **New Phytologist**, v. 214, n. 2, p. 539–553, 2017.

SHERMAN-BROYLES, S. *et al.* Complete Plastome Sequences from Glycine syndetika and Six Additional Perennial Wild Relatives of Soybean. **G3: Genes|Genomes|Genetics**, v. 4, n. 10, p. 2023–2033, 25 ago. 2014.

SIM, J. *et al.* Recent advances in the improvement of soybean seed traits by genome editing. **Plant Biotechnology**, v. 40, n. 3, p. 193–200, 25 set. 2023.

SINGH, P. *et al.* Functional and Edible Uses of Soy Protein Products. **Comprehensive Reviews in Food Science and Food Safety**, v. 7, n. 1, p. 14–28, 2008.

SONAWANE, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. **Cell reports**, v. 21, n. 4, p. 1077–1088, 24 out. 2017.

SONG, Q. *et al.* Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. **The Plant Journal**, v. 104, n. 3, p. 800, 23 set. 2020.

SOYSTATS. **Soystats**, 2025. Disponível em: <<http://soystats.com/2023-soystats/>>. Acesso em: 31 out. 2024

TETTELIN, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. **Proceedings of the National Academy of Sciences**, v. 102, n. 39, p. 13950–13955, 27 set. 2005.

TETTELIN, H.; MEDINI, D. (EDS.). **The Pangenome: Diversity, Dynamics and Evolution of Genomes**. [s.l.] Springer Nature, 2020.

USDA. **USDA**. Disponível em: <<https://www.fas.usda.gov/data/production/commodity/2222000>>. Acesso em: 9 jan. 2026.

VAN BEL, M. *et al.* PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. **Nucleic Acids Research**, v. 50, n. D1, p. D1468–D1474, 7 jan. 2022.

WANG, D. *et al.* The restorer gene for soybean M-type cytoplasmic male sterility, Rf-m, is located in a PPR gene-rich region on chromosome 16. **Plant Breeding**, v. 135, n. 3, p. 342–348, 2016.

WANG, J. *et al.* A pangenome analysis pipeline provides insights into functional gene identification in rice. **Genome Biology**, v. 24, n. 1, p. 19, 26 jan. 2023.

WANG, J.; HU, X. Research on corn production efficiency and influencing factors of typical farms: Based on data from 12 corn-producing countries from 2012 to 2019. **PLoS ONE**, v. 16, n. 7, p. e0254423, 9 jul. 2021.

WANG, M. *et al.* Genome-wide identification and expression analysis of the AT-hook Motif Nuclear Localized gene family in soybean. **BMC Genomics**, v. 22, n. 1, p. 361, 18 maio 2021.

WANG, Y. *et al.* The L-type lectin receptor-like kinase GmLecRK-IX negatively regulates drought stress responses in soybean seedlings. **Plant Physiology and Biochemistry**, v. 227, p. 110134, 1 out. 2025.

WICKHAM, H. *et al.* Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 21 nov. 2019.

WICKHAM, H.; YUTANI, H. Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 21 nov. 2019.

WU, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. **The Innovation**, v. 2, n. 3, p. 100141, 28 ago. 2021.

WU, W. *et al.* Association of extracellular dNTP utilization with a GmPAP1-like protein identified in cell wall proteomic analysis of soybean roots. **Journal of Experimental Botany**, v. 69, n. 3, p. 603–617, 23 jan. 2018.

XU, H. *et al.* Progress in Soybean Genetic Transformation Over the Last Decade. **Frontiers in Plant Science**, v. 13, 9 jun. 2022.

YANG, Z. *et al.* The analysis of the genetic loci affecting phenotypic plasticity of soybean isoflavone content by dQTG.seq model. **TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik**, v. 138, n. 1, p. 9, 17 dez. 2024.

YUAN, X. *et al.* Integrative omics analysis elucidates the genetic basis underlying seed weight and oil content in soybean. **The Plant Cell**, v. 36, n. 6, p. 2160–2175, 27 fev. 2024.

ZHANG, H. *et al.* Transcriptome profiling reveals the spatial-temporal dynamics of gene expression essential for soybean seed development. **BMC Genomics**, v. 22, n. 1, p. 453, 16 jun. 2021.

ZHAO, C. *et al.* Temperature increase reduces global yields of major crops in four independent estimates. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 35, p. 9326–9331, 29 ago. 2017.

ZHAO, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. **Nature Genetics**, v. 50, n. 2, p. 278–284, fev. 2018.

ZHOU, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. **Nature Biotechnology**, v. 33, n. 4, p. 408–414, abr. 2015.

ZHU, L. *et al.* Genome-wide identification and expression pattern analysis of the kiwifruit GRAS transcription factor family in response to salt stress. **BMC Genomics**, v. 25, n. 1, p. 12, 2 jan. 2024.

ZHUANG, Y. *et al.* Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. **Nature Plants**, v. 8, n. 3, p. 233–244, mar. 2022.

Tabela S1. Termos de Gene Ontology (GO) enriquecidos para os genes tecido-específicos nas espécies perenes do gênero *Glycine*.

GO	Descrição	Genes C/NC TS	Genes totais	Categoria	Tecido
GO:0030414	peptidase inhibitor activity	NC-TS	2	Função Biológica	Parte aérea
GO:0004857	enzyme inhibitor activity	NC-TS	2	Função Biológica	Parte aérea
GO:0004866	endopeptidase inhibitor activity	NC-TS	2	Função Biológica	Parte aérea
GO:0016835	carbon-oxygen lyase activity	NC-TS	2	Função Biológica	Parte aérea
GO:0071695	anatomical structure maturation	NC-TS	2	Processo biológico	Semente
GO:0000271	polysaccharide biosynthetic process	NC-TS	2	Processo biológico	Semente
GO:0004857	enzyme inhibitor activity	NC-TS	3	Processo biológico	Semente
GO:0009101	glycoprotein biosynthetic process	NC-TS	2	Processo biológico	Semente
GO:0010431	seed maturation	NC-TS	2	Processo biológico	Semente
GO:0015291	secondary active transmembrane transporter activity	NC-TS	2	Função Biológica	Semente
GO:0015297	antiporter activity	NC-TS	2	Processo biológico	Semente
GO:0021700	developmental maturation	NC-TS	2	Processo biológico	Semente
GO:0048609	multicellular organismal reproductive process	NC-TS	2	Processo biológico	Semente
GO:0051235	maintenance of location	NC-TS	2	Processo biológico	Semente
GO:0071215	cellular response to abscisic acid stimulus	NC-TS	2	Processo biológico	Semente
GO:0071695	anatomical structure maturation	NC-TS	2	Processo biológico	Cotilédone
GO:0010431	seed maturation	NC-TS	2	Processo biológico	Cotilédone
GO:0021700	developmental maturation	NC-TS	2	Processo biológico	Cotilédone
GO:0032504	multicellular organism reproduction	NC-TS	2	Processo biológico	Cotilédone

GO:0048609	multicellular organismal reproductive process	NC-TS	2	Processo biológico	Cotilédone
GO:0072330	monocarboxylic acid biosynthetic process	NC-TS	2	Processo biológico	Suspensor
GO:0006633	fatty acid biosynthetic process	NC-TS	2	Processo biológico	Suspensor
GO:0016853	isomerase activity	NC-TS	2	Processo biológico	Suspensor
GO:0070085	glycosylation	NC-TS	2	Processo biológico	Tegumento
GO:0006486	protein glycosylation	NC-TS	2	Processo biológico	Tegumento
GO:0009101	glycoprotein biosynthetic process	NC-TS	2	Processo biológico	Tegumento
GO:0015291	secondary active transmembrane transporter activity	NC-TS	2	Função Biológica	Tegumento
GO:0015297	antiporter activity	NC-TS	2	Função Biológica	Tegumento
GO:0043413	macromolecule glycosylation	NC-TS	2	Processo biológico	Tegumento
GO:0090501	RNA phosphodiester bond hydrolysis	NC-TS	2	Processo biológico	Endosperma
GO:0000977	RNA polymerase II transcription regulatory region sequence-specific DNA binding	NC-TS	5	Função Biológica	Endosperma
GO:0006366	transcription by RNA polymerase II	NC-TS	6	Processo biológico	Endosperma
GO:0006379	mRNA cleavage	NC-TS	2	Processo biológico	Endosperma
GO:0006869	lipid transport	NC-TS	7	Processo biológico	Endosperma
GO:0009055	electron transfer activity	NC-TS	3	Função Biológica	Endosperma
GO:0009512	cytochrome b6f complex	NC-TS	3	Componente biológico	Endosperma
GO:0010383	cell wall polysaccharide metabolic process	NC-TS	7	Processo biológico	Endosperma
GO:0010410	hemicellulose metabolic process	NC-TS	7	Processo biológico	Endosperma
GO:0010411	xyloglucan metabolic process	NC-TS	6	Processo biológico	Endosperma
GO:0010876	lipid localization	NC-TS	8	Processo biológico	Endosperma

GO:0043086	negative regulation of catalytic activity	NC-TS	2	Processo biológico	Endosperma
GO:0044042	glucan metabolic process	NC-TS	6	Processo biológico	Endosperma
GO:0046527	glucosyltransferase activity	NC-TS	6	Função Biológica	Endosperma
GO:0071369	cellular response to ethylene stimulus	NC-TS	2	Processo biológico	Vagem
GO:0000160	phosphorelay signal transduction system	NC-TS	2	Processo biológico	Vagem
GO:0003953	NAD+ nucleosidase activity	NC-TS	3	Processo biológico	Vagem
GO:0009723	response to ethylene	NC-TS	2	Processo biológico	Vagem
GO:0009873	ethylene-activated signaling pathway	NC-TS	2	Processo biológico	Vagem
GO:0016799	hydrolase activity, hydrolyzing N-glycosyl compounds	NC-TS	3	Processo biológico	Vagem
GO:0016799	hydrolase activity, hydrolyzing N-glycosyl compounds	NC-TS	7	Processo biológico	Pecíolo
GO:0003953	NAD+ nucleosidase activity	NC-TS	7	Processo biológico	Pecíolo
GO:0140359	ABC-type transporter activity	C-TS	9	Processo biológico	Raiz
GO:0000217	DNA secondary structure binding	C-TS	10	Processo biológico	Raiz
GO:0003680	minor groove of adenine-thymine-rich DNA binding	C-TS	10	Processo biológico	Raiz
GO:0004030	aldehyde dehydrogenase [NAD(P)+] activity	C-TS	2	Processo biológico	Raiz
GO:0004519	endonuclease activity	C-TS	5	Processo biológico	Raiz
GO:0004521	endoribonuclease activity	C-TS	4	Processo biológico	Raiz
GO:0004601	peroxidase activity	C-TS	33	Processo biológico	Raiz
GO:0005375	copper ion transmembrane transporter activity	C-TS	2	Processo biológico	Raiz

GO:0005381	iron ion transmembrane transporter activity	C-TS	3	Processo biológico	Raiz
GO:0005384	manganese ion transmembrane transporter activity	C-TS	3	Processo biológico	Raiz
GO:0005385	zinc ion transmembrane transporter activity	C-TS	2	Processo biológico	Raiz
GO:0005507	copper ion binding	C-TS	8	Processo biológico	Raiz
GO:0007043	cell-cell junction assembly	C-TS	10	Processo biológico	Raiz
GO:0008131	primary amine oxidase activity	C-TS	2	Processo biológico	Raiz
GO:0008171	O-methyltransferase activity	C-TS	8	Processo biológico	Raiz
GO:0008422	beta-glucosidase activity	C-TS	3	Processo biológico	Raiz
GO:0009505	plant-type cell wall	C-TS	24	Processo biológico	Raiz
GO:0009531	secondary cell wall	C-TS	15	Processo biológico	Raiz
GO:0009699	phenylpropanoid biosynthetic process	C-TS	8	Processo biológico	Raiz
GO:0009735	response to cytokinin	C-TS	15	Processo biológico	Raiz
GO:0009751	response to salicylic acid	C-TS	8	Processo biológico	Raiz
GO:0009809	lignin biosynthetic process	C-TS	6	Processo biológico	Raiz
GO:0009838	abscission	C-TS	4	Processo biológico	Raiz
GO:0009871	jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway	C-TS	2	Processo biológico	Raiz
GO:0010039	response to iron ion	C-TS	3	Processo biológico	Raiz
GO:0010227	floral organ abscission	C-TS	3	Processo biológico	Raiz
GO:0010447	response to acidic pH	C-TS	2	Processo biológico	Raiz
GO:0010930	negative regulation of auxin mediated signaling pathway	C-TS	2	Processo biológico	Raiz

GO:0015114	phosphate ion transmembrane transporter activity	C-TS	2	Processo biológico	Raiz
GO:0015144	carbohydrate transmembrane transporter activity	C-TS	5	Processo biológico	Raiz
GO:0015399	primary active transmembrane transporter activity	C-TS	10	Processo biológico	Raiz
GO:0015926	glucosidase activity	C-TS	3	Processo biológico	Raiz
GO:0015928	fucosidase activity	C-TS	2	Processo biológico	Raiz
GO:0016145	S-glycoside catabolic process	C-TS	2	Processo biológico	Raiz
GO:0016209	antioxidant activity	C-TS	33	Processo biológico	Raiz
GO:0016679	oxidoreductase activity, acting on diphenols and related substances as donors	C-TS	6	Processo biológico	Raiz
GO:0016682	oxidoreductase activity, acting on diphenols and related substances as donors, oxygen as acceptor	C-TS	6	Processo biológico	Raiz
GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	C-TS	33	Processo biológico	Raiz
GO:0016849	phosphorus-oxygen lyase activity	C-TS	4	Processo biológico	Raiz
GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	C-TS	5	Processo biológico	Raiz
GO:0019137	thioglucosidase activity	C-TS	2	Processo biológico	Raiz
GO:0019759	glycosinolate catabolic process	C-TS	2	Processo biológico	Raiz
GO:0019762	glucosinolate catabolic process	C-TS	2	Processo biológico	Raiz
GO:0030145	manganese ion binding	C-TS	7	Processo biológico	Raiz
GO:0031012	extracellular matrix	C-TS	3	Processo biológico	Raiz
GO:0033897	ribonuclease T2 activity	C-TS	4	Processo biológico	Raiz
GO:0034329	cell junction assembly	C-TS	10	Processo biológico	Raiz

GO:0034330	cell junction organization	C-TS	10	Processo biológico	Raiz
GO:0042545	cell wall modification	C-TS	17	Processo biológico	Raiz
GO:0042626	ATPase-coupled transmembrane transporter activity	C-TS	10	Processo biológico	Raiz
GO:0042744	hydrogen peroxide catabolic process	C-TS	33	Processo biológico	Raiz
GO:0042803	protein homodimerization activity	C-TS	10	Processo biológico	Raiz
GO:0044550	secondary metabolite biosynthetic process	C-TS	9	Processo biológico	Raiz
GO:0045216	cell-cell junction organization	C-TS	10	Processo biológico	Raiz
GO:0045841	negative regulation of mitotic metaphase/anaphase transition	C-TS	2	Processo biológico	Raiz
GO:0046271	phenylpropanoid catabolic process	C-TS	5	Processo biológico	Raiz
GO:0046274	lignin catabolic process	C-TS	5	Processo biológico	Raiz
GO:0046864	isoprenoid transport	C-TS	2	Processo biológico	Raiz
GO:0046865	terpenoid transport	C-TS	2	Processo biológico	Raiz
GO:0046915	transition metal ion transmembrane transporter activity	C-TS	5	Processo biológico	Raiz
GO:0048226	Casparian strip	C-TS	15	Processo biológico	Raiz
GO:0048510	regulation of timing of transition from vegetative to reproductive phase	C-TS	4	Processo biológico	Raiz
GO:0051119	sugar transmembrane transporter activity	C-TS	4	Processo biológico	Raiz
GO:0052716	hydroquinone:oxygen oxidoreductase activity	C-TS	5	Processo biológico	Raiz
GO:0072593	reactive oxygen species metabolic process	C-TS	34	Processo biológico	Raiz
GO:0072686	mitotic spindle	C-TS	2	Processo biológico	Raiz
GO:0097237	cellular response to toxic substance	C-TS	3	Processo biológico	Raiz

Tabela S2. Termos de Gene Ontology (GO) enriquecidos para os genes tecido-específicos em *G. soja*.

GO	Descrição	Genes C/NC TS	Genes totais	Categoria	Tecido
GO:0051235	maintenance of location	NC	1	Processo Biológico	Semente
GO:0005811	lipid droplet	NC	1	Componente Biológico	Semente
GO:0006636	unsaturated fatty acid biosynthetic process	NC	1	Processo Biológico	Semente
GO:0006829	zinc ion transport	NC	1	Processo Biológico	Semente
GO:0009269	response to desiccation	NC	1	Processo Biológico	Semente
GO:0012511	monolayer-surrounded lipid storage body	NC	1	Componente Biológico	Semente
GO:0019915	lipid storage	NC	1	Processo Biológico	Semente
GO:0031408	oxylipin biosynthetic process	NC	1	Processo Biológico	Semente
GO:0034389	lipid droplet organization	NC	1	Processo Biológico	Semente
GO:0042759	long-chain fatty acid biosynthetic process	NC	1	Processo Biológico	Semente
GO:2000377	regulation of reactive oxygen species metabolic process	NC	1	Processo Biológico	Cotilédone
GO:0005811	regulation of reactive oxygen species metabolic process	NC	1	Processo Biológico	Cotilédone
GO:0006636	unsaturated fatty acid biosynthetic process	NC	1	Processo Biológico	Cotilédone
GO:0009269	response to desiccation	NC	1	Processo Biológico	Cotilédone
GO:0012511	monolayer-surrounded lipid storage body	NC	1	Componente Biológico	Cotilédone
GO:0019915	lipid storage	NC	1	Processo Biológico	Cotilédone
GO:0031408	oxylipin biosynthetic process	NC	1	Processo Biológico	Cotilédone
GO:0034389	lipid droplet organization	NC	1	Processo Biológico	Cotilédone

GO:0042759	long-chain fatty acid biosynthetic process	NC	1	Processo Biológico	Cotilédone
GO:0051235	maintenance of location	NC	1	Processo Biológico	Cotilédone
GO:0070069	cytochrome complex	NC	2	Componente Biológico	Suspensor
GO:0001676	long-chain fatty acid metabolic process	NC	1	Processo Biológico	Suspensor
GO:0006099	tricarboxylic acid cycle	NC	4	Processo Biológico	Suspensor
GO:0006633	fatty acid biosynthetic process	NC	2	Processo Biológico	Suspensor
GO:0008964	phosphoenolpyruvate carboxylase activity	NC	4	Função Biológica	Suspensor
GO:0009055	electron transfer activity	NC	2	Função Biológica	Suspensor
GO:0009512	cytochrome b6f complex	NC	2	Componente Biológico	Suspensor
GO:0015977	carbon fixation	NC	4	Processo Biológico	Suspensor
GO:0031408	oxylipin biosynthetic process	NC	1	Processo Biológico	Suspensor
GO:0042759	long-chain fatty acid biosynthetic process	NC	1	Processo Biológico	Suspensor
GO:0030246	carbohydrate binding	NC	1	Função Biológica	Tegumento
GO:0030117	membrane coat	NC	1	Componente Biológico	Tegumento
GO:0070069	cytochrome complex	NC	2	Componente Biológico	Endosperma
GO:0000105	histidine biosynthetic process	NC	1	Processo Biológico	Endosperma
GO:0001676	long-chain fatty acid metabolic process	NC	1	Processo Biológico	Endosperma
GO:0006099	tricarboxylic acid cycle	NC	6	Processo Biológico	Endosperma
GO:0008964	phosphoenolpyruvate carboxylase activity	NC	6	Função Biológica	Endosperma
GO:0009055	electron transfer activity	NC	2	Função Biológica	Endosperma
GO:0009512	cytochrome b6f complex	NC	2	Componente Biológico	Endosperma
GO:0015977	carbon fixation	NC	6	Processo Biológico	Endosperma

GO:0031408	oxylipin biosynthetic process	NC	1	Processo Biológico	Endosperma
GO:0042759	long-chain fatty acid biosynthetic process	NC	1	Processo Biológico	Endosperma
GO:0140546	defense response to symbiont	NC	1	Processo Biológico	Parte aérea
GO:0003953	NAD+ nucleosidase activity	NC	1	Função Biológica	Parte aérea
GO:0004252	serine-type endopeptidase activity	NC	1	Função Biológica	Parte aérea
GO:0006298	mismatch repair	NC	1	Processo Biológico	Parte aérea
GO:0006633	fatty acid biosynthetic process	NC	1	Processo Biológico	Parte aérea
GO:0008081	phosphoric diester hydrolase activity	NC	1	Processo Biológico	Parte aérea
GO:0016835	carbon-oxygen lyase activity	NC	1	Função Biológica	Parte aérea
GO:0016836	hydro-lyase activity	NC	1	Função Biológica	Parte aérea
GO:0030983	mismatched DNA binding	NC	1	Função Biológica	Parte aérea
GO:0031047	gene silencing by RNA	NC	1	Processo Biológico	Parte aérea
GO:0042759	long-chain fatty acid biosynthetic process	NC	1	Processo Biológico	Radícula
GO:0006636	unsaturated fatty acid biosynthetic process	NC	1	Processo Biológico	Radícula
GO:0006829	zinc ion transport	NC	1	Processo Biológico	Radícula
GO:0009512	cytochrome b6f complex	NC	1	Componente Biológico	Radícula
GO:0012511	monolayer-surrounded lipid storage body	NC	1	Componente Biológico	Radícula
GO:0016886	ligase activity, forming phosphoric ester bonds	NC	1	Função Biológica	Radícula
GO:0017004	cytochrome complex assembly	NC	1	Processo Biológico	Radícula
GO:0019915	lipid storage	NC	1	Processo Biológico	Radícula
GO:0031408	oxylipin biosynthetic process	NC	1	Processo Biológico	Radícula
GO:0034389	lipid droplet organization	NC	1	Processo Biológico	Radícula

GO:0048226	Casparian strip	C	15	Componente Biológico	Raiz
GO:0000217	DNA secondary structure binding	C	10	Função Biológica	Raiz
GO:0003680	minor groove of adenine-thymine-rich DNA binding	C	10	Função Biológica	Raiz
GO:0004601	peroxidase activity	C	32	Função Biológica	Raiz
GO:0009505	plant-type cell wall	C	23	Componente Biológico	Raiz
GO:0009698	phenylpropanoid metabolic process	C	17	Processo Biológico	Raiz
GO:0034330	cell junction organization	C	10	Processo Biológico	Raiz
GO:0042545	cell wall modification	C	17	Processo Biológico	Raiz
GO:0042744	hydrogen peroxide catabolic process	C	32	Processo Biológico	Raiz
GO:0045216	cell-cell junction organization	C	10	Processo Biológico	Raiz
GO:0052689	carboxylic ester hydrolase activity	C	19	Função Biológica	Flor
GO:0000977	RNA polymerase II transcription regulatory region sequence-specific DNA binding	C	33	Função Biológica	Flor
GO:0004857	enzyme inhibitor activity	C	32	Função Biológica	Flor
GO:0010093	specification of floral organ identity	C	9	Processo Biológico	Flor
GO:0010584	pollen exine formation	C	12	Processo Biológico	Flor
GO:0030599	pectinesterase activity	C	19	Função Biológica	Flor
GO:0032989	cellular component morphogenesis	C	13	Processo Biológico	Flor
GO:0042545	cell wall modification	C	23	Processo Biológico	Flor
GO:0048438	floral whorl development	C	23	Processo Biológico	Flor
GO:0048646	anatomical structure formation involved in morphogenesis	C	24	Processo Biológico	Flor
GO:0042744	hydrogen peroxide catabolic process	C	17	Processo Biológico	Nódulo
GO:0003680	minor groove of adenine-thymine-rich DNA binding	C	8	Função Biológica	Nódulo

GO:0004601	peroxidase activity	C	18	Função Biológica	Nódulo
GO:0008131	primary amine oxidase activity	C	5	Função Biológica	Nódulo
GO:0008422	beta-glucosidase activity	C	6	Função Biológica	Nódulo
GO:0009735	response to cytokinin	C	15	Processo Biológico	Nódulo
GO:0015926	glucosidase activity	C	7	Função Biológica	Nódulo
GO:0016709	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen	C	8	Função Biológica	Nódulo
GO:0019137	thioglucosidase activity	C	4	Função Biológica	Nódulo
GO:0019825	oxygen binding	C	5	Função Biológica	Nódulo

Tabela S3. Ortogrupos inspecionados manualmente para prospecção de genes de interesse.

Ortogrupo	Função GO principal	Tecido TS	G. max	G. soja	Perenes	Potencial aplicado
Glycine.pan5.pan55826	EF-hand domain-containing protein	Cotilédone / Vagem / Semente / Tegumento / Suspensor / Endosperma / Radícula / Raiz	1	6	1	resposta a estresse e desenvolvimento de semente
Glycine.pan5.pan44341	GH16 domain-containing protein	Endosperma	1	0	1	Remodelação de parede celular e desenvolvimento
Glycine.pan5.pan61024	GH16 domain-containing protein	Endosperma	1	0	1	Remodelação de parede celular e desenvolvimento
Glycine.pan5.pan46593	ADP-ribosyl cyclase/ cyclic ADP-ribose hydrolase	Peciolo	1	0	1	Sinalização por cADPR/Ca2+; respostas a estresses
Glycine.pan5.pan38640	ADP-ribosyl cyclase/ cyclic ADP-ribose hydrolase	Peciolo	1	0	1	Sinalização por cADPR/Ca2+; respostas a estresses
Glycine.pan5.pan48839	TIR domain-containing protein	Peciolo	1	0	1	resistência a patógenos
Glycine.pan5.pan46867	TIR domain-containing protein	Peciolo	1	0	1	resistência a patógenos
Glycine.pan5.pan50835	TIR domain-containing protein/ Candidate disease-resistance protein	Peciolo	1	0	1	resistência a patógenos
Glycine.pan5.pan50837	R protein	Peciolo	1	0	1	resistência a patógenos
Glycine.pan5.pan59302	NB-ARC domain-containing protein	Peciolo	1	0	1	resistência a patógenos
Glycine.pan5.pan01292	Peroxidase	Raiz / Nódulo	1	2	1	Homeostase redox e lignificação; tolerância a estresses e defesa
Glycine.pan5.pan28265	CASP-like protein/ Casparian strip membrane protein 1	Raiz	1	0	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes

Glycine.pan5.pan28286	Laccase	Raiz	1	0	1	Lignificação/parede celular; resistência estrutural e adaptação a estresses
Glycine.pan5.pan29868	CASP-like protein	Raiz	1	1	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes
Glycine.pan5.pan29869	CASP-like protein	Raiz	1	1	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes
Glycine.pan5.pan30356	CASP-like protein/ Casparian strip membrane protein 5	Raiz	1	1	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes
Glycine.pan5.pan14529	CASP-like protein	Raiz	1	1	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes
Glycine.pan5.pan15484	Laccase	Raiz	1	0	1	Lignificação/parede celular; resistência estrutural e adaptação a estresses
Glycine.pan5.pan19613	CASP-like protein	Raiz	1	0	1	Faixa de Caspary/barreira radicular; eficiência hídrica e de nutrientes
Glycine.pan5.pan36674	Bet v I/ Major latex protein domain-containing protein	Semente / Tegumento	1	0	2	Proteína MLP/defesa; tolerância a estresse em sementes
Glycine.pan5.pan17110	Fucosyltransferase	Semente / Tegumento	1	0	2	Modificação de carboidratos; desenvolvimento/qualidade de semente
Glycine.pan5.pan01014	Uncharacterized protein	Plântula (aérea)	1	0	1	Gene candidato para caracterização; possível variante adaptativa
Glycine.pan5.pan62517	Codifica la subunidad 8 del complejo del citocromo b6-f	Suspensor / Endosperma / Radícula	1	3	0	Fotossíntese (citocromo b6f); eficiência energética e produtividade

Glycine.pan5.pan00041	Peroxidase 39-like	Raiz / Nódulo	1	2	0	Homeostase redox e lignificação; tolerância a estresses e defesa
Glycine.pan5.pan31215	Peroxidase 59	Raiz / Nódulo	1	2	0	Homeostase redox e lignificação; tolerância a estresses e defesa
Glycine.pan5.pan26705	Peroxidase 59	Raiz / Nódulo	1	2	0	Homeostase redox e lignificação; tolerância a estresses e defesa
