

**UNIVERSIDADE FEDERAL DE MINAS GERAIS**  
Instituto de Ciências Biológicas  
Programa Interunidades de Pós-Graduação em Bioinformática

Mariana Costa Dias

**ADAPTAÇÕES MOLECULARES NO DESENVOLVIMENTO DE *PARKIA*  
*PLATYCEPHALA* E *STRYPHNODEDRON PULCHERRIMUM* EM SUBSTRATOS  
DE CANGA E RESÍDUOS DE MINERAÇÃO**

Belo Horizonte

2021

Mariana Costa Dias

**ADAPTAÇÕES MOLECULARES NO DESENVOLVIMENTO DE *PARKIA*  
*PLATYCEPHALA* E *STRYPHODENDRON PULCHERRIMUM* EM SUBSTRATOS  
DE CANGA E RESÍDUOS DE MINERAÇÃO**

**Versão final**

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Bioinformática.

Orientador: Dr. Guilherme Oliveira

Belo Horizonte

2021

043

Dias, Mariana Costa.

Adaptações moleculares no desenvolvimento de *Parkia platycephala* e *Stryphnodendron pulcherrimum* em substratos de canga e resíduos de mineração [manuscrito] / Mariana Costa Dias. – 2021.

110 f. : il. ; 29,5 cm.

Orientador: Dr. Guilherme Corrêa de Oliveira.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. Savana. 3. Mineração. 4. Recuperação e Remediação Ambiental. 5. Fabaceae. 6. Expressão Gênica. 7. Adaptação Fisiológica. 8. Estresse Fisiológico. 9. Transcriptoma. I. Oliveira, Guilherme Corrêa de. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS BIOLÓGICAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

**PARECER Nº**  
PROCESSO Nº

**36/2021**  
23072.263921/2021-76

**FOLHA DE APROVAÇÃO**

**"Adaptações moleculares no desenvolvimento de *Parkia platycephala* e *Stryphnodendron pulcherrimum* em substratos de canga e resíduos de mineração"**

**Mariana Costa Dias**

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Guilherme Corrêa de Oliveira - Orientador  
Instituto Tecnológico Vale

Profa Valesca Pandolfi  
Universidade Federal de Pernambuco

Prof. Aristóteles Góes Neto  
Universidade Federal de Minas Gerais

Prof. Rommel Thiago Juca Ramos  
Universidade Federal do Pará

Profa Priscila Grynberg  
Embrapa

Belo Horizonte, 15 de dezembro de 2021.



Documento assinado eletronicamente por **Guilherme Corrêa de Oliveira, Usuário Externo**, em 15/12/2021, às 16:45, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Valesca Pandolfi, Usuário Externo**, em 15/12/2021, às 16:49, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Rommel Thiago Jucá Ramos, Usuário Externo**, em 15/12/2021, às 16:50, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Priscila Grynberg, Usuária Externa**, em 15/12/2021, às 16:51, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Aristoteles Goes Neto, Coordenador(a) de curso de pós-graduação**, em 15/12/2021, às 16:57, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site [https://sei.ufmg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **1146012** e o código CRC **54746031**.

Todas as conquistas são e serão sempre dedicadas a  
ela, minha mãe, Ana Lúcia.

## AGRADECIMENTOS

Primeiramente gostaria de agradecer ao meu orientador, Dr. Guilherme Corrêa de Oliveira. Guilherme me recebeu em Belém, no ITV e na sua casa, de braços abertos, muita empolgação e muitas ideias na cabeça ainda antes de eu decidir entrar para o Doutorado. Depois que me decidi, me acolheu na escolha do programa, confiou em mim e possibilitou meu crescimento na Bioinformática, uma área que eu nunca tinha trabalhado antes. Já são cinco anos de interação e serei eternamente grata por todo o aprendizado e todas as oportunidades que tive com ele e no ITV. Mudar para Belém e trabalhar no ITV foram duas das melhores escolhas que já fiz e muito disso se deve a ele. Muito obrigada!

Falando de Belém, não posso deixar de agradecer aos grandes amigos que fiz nessa cidade incrível! À Gigege, por todo o acolhimento com que me recebeu em Belém, por dividir comigo milhões de experiências massa em todos os lugares que me levou, por ser a melhor gestora (ainda sem ser)! Por todas as risadas e as raivas que passamos juntas, por todas as comidas e botecos, por toda essa parceria. Ao Santelmo, a pessoa que me treinou no laboratório do ITV e um dos biólogos que eu mais admiro. Obrigada por dividir comigo seu conhecimento, sua bancada, suas pipetas (que ninguém podia pegar, mas... \o/), sua rabugice, paciência, suas terças, sua casa, por todo o carinho e confiança. Ao Renato, meu amigo, colega de Doutorado e de orientador, obrigada por estar lá sempre que precisei, por todo suporte, apoio, pelas dicas e scripts. Obrigada por jogar no meu time sempre, me entender e me enxergar. À Sílvia, que se tornou uma amiga muito especial e querida. Obrigada por todas as conversas, todo o carinho e consideração e toda a torcida. Ao Rafa, a primeira pessoa desse grupo a me incentivar a fazer Doutorado, o que mais acreditou, o que mais “martelou na minha cabeça”. Eu senti e sinto essa fé por todo o processo, em cada apresentação e conversa. E em especial, ao Eder, que viveu comigo todo o meu tempo de Belém, todas as alegrias e todos os apereios. Viveu todo o processo do Doutorado, desde a prova de seleção e escrita do projeto, o estresse da qualificação. Dividiu todas as alegrias e as angústias do ITV. Obrigada por me acompanhar nessa jornada. Você é minha casa e tem uma casa em mim também. Mais que um amigo, eu ganhei um irmão. Cada um da sua forma, vocês foram extremamente importantes nessa etapa da minha vida. Eu sou muito feliz em ter o carinho, o reconhecimento e a amizade de cada um de vocês!

A todos os amigos que passaram pelo ITV e pelo grupo de Genômica Ambiental antigos e novos, em especial Talvâne, Camila, Marx e Michele. Ao Manoel, que é quem faz tudo funcionar no laboratório! Por toda sua dedicação e bom humor, por toda a ajuda e carinho. Ao

Fabrcio, responsvel pelo cluster do ITV (devia ter uma pgina inteira de agradecimento s para ele!), por todas as vezes que me salvou: quando o cluster travava, quando precisava de qualquer programa instalado, quando ficamos sem acesso e te fiz tirar *print* de quase todos os caminhos de diretrios para recuperar dados... a lista vai alm! Muito obrigada por toda ajuda e disponibilidade!

Ao grupo de Recuperao de reas Degradadas do ITV pela colaborao e fornecimento das amostras utilizadas nesse trabalho. Em especial ao Cecilio pelas valiosas contribuies, por sempre se disponibilizar a ajudar e pela amizade.

A todos os amigos do LGB e da Bio, principalmente a Mika, Mari Babys, Lili, Thomaz, Pepelli, Stellinha, Cae e Ello. A Nayara, amiga e colega de Doutorado, que dividiu angustias e calmarias, a maior portadora de boas notcias! Ao Repolês, uma das pessoas mais importantes de todas as minhas etapas acadêmicas at agora, acho que nunca agradei o suficiente! Obrigada por todos os conselhos, por todas as discusses, cientficas ou no, por estar sempre comigo, torcer e acreditar em mim. Passar por uma pandemia nos ltimos anos de Doutorado no foi fcil (nem pra mim, nem pra ningum) e mais uma vez você foi essencial. Obrigada por perguntar, mesmo quando eu parei de responder, e por continuar perguntando.

Aos amigos do Eixo, que me acompanham desde o colgio. Em especial, Marininha, Luiza e Foca, que me acolhem, e cada um do seu jeito, me fazem enxergar as coisas de outras formas e me relembram constantemente o valor das amizades. A saudade de vocs no cabe em mim!

A minha famlia, que me ajuda construir meu castelinho. As minhas primas, Dani, Bruna, Cacá, Lu, Bela e Nina. Vocs foram minhas companheiras de aventuras por toda a minha vida desde sempre. Foi principalmente com vocs que eu tive meus maiores vacilos e acertos, onde eu briguei, chorei, ri e amei mais. Crescer com vocs me ensinou a ser quem eu sou hoje. Muito obrigada! Aos meus irmos, Bê e Nanda, os mais doces e carinhosos! Vocs inspiram afeto e ternura onde passam e comigo no seria diferente! Obrigada pela torcida e por todo o amor que vocs emanam. Ao meu pai, Sandro, por todo apoio e amor incondicional. Em especial, a minha me, Ana Lcia. Minha referncia de coragem e fora. Sem ela nada disso seria possvel (clichê é clichê por ser verdade absoluta – outro clichê). Obrigada por todo amor e dedicao, por todo o incentivo e apoio. Você é minha base! E eu tenho a base mais forte de todas!

A Dona Nilde, Teta e Daniella, o carinho com que vocs me receberam na famlia de vocs é inexplicvel! Obrigada por me receberem sempre com tanta leveza, tanto amor e tanta mangao! Eu tenho a melhor sogra e as melhores cunhadas!

Ao Diego, meu parceiro de vida, meu amor e meu melhor amigo. Obrigada por tornar essa jornada mais leve e divertida. Obrigada por todos os conselhos e por tanta cumplicidade. Você me mostra o real significado de companheirismo todos os dias, nas pequenas coisas. Você cuida de mim, dos meus, da casa, traz minha água e fez tudo ao seu alcance para que o término dessa fase fosse mais tranquilo e feliz. Seu entusiasmo pela ciência e por dar aulas me inspiram a ser uma pesquisadora muito melhor. Muito obrigada pelo amor, carinho e compreensão. Você extrai o melhor de mim.

Agradeço, por fim, à Vale pela criação e financiamento do Instituto Tecnológico Vale e pelo investimento no projeto, e ao CNPq e à FAPESP pela concessão das bolsas de estudo em parceria com o ITV.

Muito obrigada!

"Science is, after all, the abstraction of simple principles from complex data."

Mensh & Kording

## RESUMO

Canga é o termo brasileiro utilizado para vegetação tipo savana que abriga várias espécies endêmicas localizadas em afloramentos rochosos ricos em ferro, frequentemente considerados para atividades de mineração. A legislação exige a reabilitação de áreas degradadas para minimizar os impactos ambientais. Apesar da urgente demanda por mineração sustentável, a revegetação e reabilitação de áreas degradadas pelas atividades de mineração permanecem desafiadoras. *Parkia platycephala* Benth. e *Stryphnodendron pulcherrimum* (Willd.) Hochr. ocorrem naturalmente nas cangas da Serra dos Carajás (leste da Amazônia, Pará, Brasil) e na floresta circundante, indicando elevada plasticidade fenotípica, e vêm sendo estudadas para recuperação de áreas de mineração. Os mecanismos morfológicos e fisiológicos do estabelecimento das plantas no ambiente da canga são bem estudados, mas as respostas moleculares adaptativas ainda não foram desvendadas. Nesse trabalho, os transcriptomas do tecido foliar de duas espécies de Fabaceae foram gerados e usados para caracterizar a plasticidade da expressão gênica e compreender os mecanismos genéticos adaptativos por trás do estabelecimento dessas plantas nos ambientes de canga e floresta. Além disso, examinamos o crescimento das espécies em substratos de resíduos de mineração para avaliar se a modulação da expressão gênica promove adaptação a esses substratos. Os resultados revelam mecanismos moleculares relevantes para a reabilitação de áreas, especialmente sob um clima em mudança, além de prever o sucesso no estabelecimento das espécies. As plantas foram cultivadas em substratos de canga, mata e resíduos de mina coletados na Província Mineral de Carajás, Pará, Brasil. O RNA das folhas foi extraído e as bibliotecas produzidas foram sequenciadas na plataforma Illumina com sequências do tipo *paired-end* de 150 pares de base. Uma combinação de métodos foi utilizada para montar as *reads* geradas nos transcriptomas de referência de *P. platycephala* e *S. pulcherrimum* contendo 31.728 e 31.311 transcritos primários, respectivamente. Usando análises de expressão diferencial, identificamos respostas moleculares compartilhadas e específicas de cada espécie para o crescimento das plantas nos substratos de canga e mata, e nos substratos de estéril de mina. Nas análises espécie-específicas nos substratos de ocorrência natural, identificamos 1.112 e 838 genes diferencialmente expressos para *P. platycephala* e *S. pulcherrimum*, respectivamente. As análises de enriquecimento mostraram processos biológicos e vias metabólicas únicas alteradas para cada espécie. Ortólogos de cópia única compartilhados foram usados para a análise comparativa de expressão diferencial. Embora o padrão geral da expressão dos ortólogos seja espécie-específico, quase 300 genes foram identificados como diferencialmente expressos entre canga

e mata respondendo da mesma forma nas espécies. Esses genes são funcionalmente associados à resposta a estímulos luminosos e à via do ritmo circadiano. O desempenho das espécies nos substratos de áreas degradadas também foi avaliado, indicando diferentes performances no desenvolvimento relacionadas às diferentes necessidades nutricionais entre elas. *S. pulcherrimum* apresentou respostas adaptativas relacionadas à limitação de fósforo nos substratos. *P. platycephala* teve mais sucesso no estabelecimento nos substratos de estéril, não apresentando alterações relacionadas a condições de estresse. Os genes ortólogos do ritmo circadiano também foram alterados de maneira similar nas duas espécies na resposta aos substratos de estéril. Nossos resultados sugerem que essas plantas são capazes de ajustar o ritmo circadiano de forma dependente do substrato. A modulação gênica do relógio circadiano pode ser um mecanismo central que regula o desenvolvimento das plantas de acordo com a disponibilidade de nutrientes nas espécies de leguminosas estudadas e pode ser compartilhado como um mecanismo comum para a compensação do estresse abiótico em outras espécies nativas. Este estudo mostra que as plantas estudadas possuem mecanismos comuns e específicos que viabilizam sua sobrevivência em ambientes tão distintos como a floresta e a canga. Os resultados contribuem para a melhoria de práticas de recuperação de áreas degradadas e a biologia de Fabaceae, especialmente em um cenário de mudanças climáticas.

Palavras-chave: *Caesalpinioideae*. Campos ferruginosos. Estresse abiótico. *Fabaceae*. Plasticidade da expressão gênica. RNA-seq. Recuperação de áreas degradadas. Transcriptoma *de novo*.

## ABSTRACT

*Canga* is the Brazilian term for savanna-like vegetation harboring several endemic species on iron-rich rocky outcrops and often industrial mining activities. Environmental legislation demands the rehabilitation of degraded areas to minimize environmental impacts. Despite the urgent demand for sustainable mining, the revegetation and rehabilitation of areas degraded by mining activities remain challenging. *Parkia platycephala* Benth. and *Stryphnodendron pulcherrimum* (Willd.) Hochr. naturally occur in the *cangas* of Serra dos Carajás (eastern Amazonia, Brazil) and the surrounding forest, indicating high phenotypic plasticity, and have been studied for the recovery of mining areas. The morphological and physiological mechanisms of the plants' establishment in the *canga* environment are well studied, but not the molecular adaptive responses. We constructed complete and accurate transcriptomes from the two Fabaceae species to characterize the gene expression plasticity and understand the adaptive genetic mechanisms behind the establishment in the *canga* and forest environments. Further, we examined the species growth in mine wastes to evaluate if the shift in the gene expression profile fosters acclimation. The results may reveal relevant molecular mechanisms for land rehabilitation, especially under a changing climate. Plants were grown in substrates of *canga*, an adjacent forest, and mine residues collected in the Carajás Mineral Province, Pará, Brazil. RNA was extracted from pooled leaf tissue. A combination of methods was utilized to assemble Illumina RNA-seq paired-end reads into two representative transcriptomes for *P. platycephala* and *S. pulcherrimum* containing 31,728 and 31,311 primary transcripts, respectively. We identified shared and species-specific molecular responses to the plants' growth in the substrates using differential expression analysis. In the species-specific analysis on naturally occurring substrates, we identified 1,112 and 838 differentially expressed genes for *P. platycephala* and *S. pulcherrimum*, respectively. Enrichment analyses showed unique biological processes and metabolic pathways affected for each species. Comparative differential expression analysis was based on shared single-copy orthologs. The overall pattern of ortholog expression was species-specific. Even so, almost 300 genes were identified differentially expressed between plants in *canga* and forest substrates, responding the same way in both species. The genes were functionally associated with the response to light stimulus and the circadian rhythm pathway. We also evaluate the performance of the species on degraded area substrates, indicating different performances in the development related to different nutritional needs among them. *S. pulcherrimum* showed adaptive responses related to phosphorus limitation in substrates. *P. platycephala* was more successful in the mine waste

substrates, showing no changes related to stress conditions. Orthologous circadian rhythm genes were also similarly changed in both species in response to the mine waste substrates. Our results suggest that plants can adjust the circadian rhythm in a substrate-dependent manner. The circadian clock gene modulation might be a central mechanism regulating the plants' development according to nutrient availability in the studied legume species, shared as a common mechanism to abiotic stress compensation in other native species. This study demonstrates that the plants have common and specific mechanisms that enable their survival in distinct environments such as forest and *canga*. The results contribute to the enhancement of land rehabilitation practices and the biology of Fabaceae, especially in a climate-changing scenario.

Keywords: Abiotic stress. *Caesalpinioideae*. Comparative transcriptomics. *De novo* transcriptome. *Fabaceae*. Gene expression plasticity. Ironstone outcrops. *Leguminosae*. Recovery of degraded areas. RNA-seq.

## LISTA DE FIGURAS

<b>Figura 1:</b> Fitofisionomias associados às cangas da Serra dos Carajás. ....	19
<b>Figura 2:</b> Mapa evidenciando o local de coleta dos diferentes substratos utilizados no experimento. ....	27
<b>Figura 3:</b> Representação da estratégia de montagem dos transcriptomas de referência. ....	33
<b>Figura 4:</b> Representação dos diferentes tipos de análises de expressão diferencial utilizados nesse trabalho. ....	38
<b>Figura 5:</b> Eletroferogramas das amostras de RNA de <i>P. platycephala</i> (cima) e <i>S. pulcherrimum</i> (baixo) evidenciando os picos 25S e 18S. ....	41
<b>Figura 6:</b> Perfil eletroforético virtual das bibliotecas de cDNA evidenciando o tamanho médio dos fragmentos obtidos. ....	42
<b>Figura 7:</b> Árvore de máxima verossimilhança aproximada baseada em ortólogos de cópia única de espécies de Caesalpinioideae, gerada com OrthoFinder/FastTree. ....	50
<b>Figura 8:</b> Perfil de expressão gênica de <i>P. platycephala</i> e <i>S. pulcherrimum</i> no desenvolvimento nos substratos naturais. ....	52
<b>Figura 9:</b> 15 principais processos biológicos enriquecidos ( $p < 0,01$ ) nos GDEs para (a) <i>P. platycephala</i> e (b) <i>S. pulcherrimum</i> . ....	54
<b>Figura 10:</b> Gráfico de dispersão das vias KEGG enriquecidas em cada espécie cultivadas em substratos de canga em comparação com substratos de mata. ....	55
<b>Figura 11:</b> Perfil de expressão gênica de <i>P. platycephala</i> , <i>S. pulcherrimum</i> no desenvolvimento nos substratos de canga e resíduos de mineração. ....	58
<b>Figura 12:</b> Gráfico de dispersão das vias KEGG enriquecidas em cada espécie cultivadas em substratos de estéril de mina em comparação com substrato de canga. ....	61
<b>Figura 13:</b> Processos biológicos enriquecidos (FDR $< 0,05$ ) nos GDEs para <i>P. platycephala</i> cultivada em estéril amarelo. ....	62
<b>Figura 14:</b> Processos biológicos enriquecidos (FDR $< 0,05$ ) nos GDEs para <i>S. pulcherrimum</i> cultivada em (a) estéril vermelho e (b) estéril amarelo. ....	63

## LISTA DE TABELAS

<b>Tabela 1:</b> Atributos químicos e físicos dos substratos coletados.....	28
<b>Tabela 2:</b> Espécies utilizadas na reconstrução da árvore filogenética.....	36
<b>Tabela 3:</b> Concentração e integridade dos RNAs extraídos. ....	42
<b>Tabela 4:</b> Concentração e tamanho médio de fragmentos das bibliotecas de cDNA. ....	43
<b>Tabela 5:</b> Resultado do sequenciamento das bibliotecas de cDNA.....	43
<b>Tabela 6:</b> Resultado do processamento das reads por qualidade e remoção de rRNA. ....	44
<b>Tabela 7:</b> Resumo das montagens.....	47
<b>Tabela 8:</b> Resultados do BUSCO com os transcriptomas de novo de <i>P. platycephala</i> e <i>S. pulcherrimum</i> . ....	49
<b>Tabela 9:</b> Plasticidade da expressão gênica das plantas no desenvolvimento em substratos de canga e mata para todos os genes e para os pares de genes ortólogos de <i>P. platycephala</i> e <i>S. pulcherrimum</i> . ....	53
<b>Tabela 10:</b> Termos Gene Ontology (GO) identificados como enriquecidos pelo teste exato de Fisher entre os ODEsRS nas plantas cultivadas na canga. ....	57
<b>Tabela 11:</b> Número de genes diferencialmente expressos no desenvolvimento em substratos de estéril de mina considerando todos os genes das espécies e para os pares de genes ortólogos de <i>P. platycephala</i> e <i>S. pulcherrimum</i> .....	60
<b>Tabela 12:</b> Termos Gene Ontology (GO) identificados como enriquecidos pelo teste exato de Fisher entre os ODEsRS nas plantas cultivadas em estéril amarelo.....	64
<b>Tabela 13:</b> Genes ortólogos diferencialmente regulados entre as espécies na resposta ao substrato de estéril amarelo. ....	66

# SUMÁRIO

<b>1. INTRODUÇÃO</b>	<b>18</b>
1.1 A canga	18
1.2 Mineração e a reabilitação das áreas degradadas	20
1.3 Plasticidade da expressão gênica e a seleção de espécies para recuperação	21
1.4 Plantas selecionadas	23
<b>2. OBJETIVOS</b>	<b>26</b>
2.1 Objetivo geral	26
2.2 Objetivos específicos	26
<b>3. MATERIAL E MÉTODOS</b>	<b>27</b>
3.1 Material biológico e desenho experimental	27
3.2 Extração de RNA	29
3.3 Preparo das bibliotecas e sequenciamento	29
3.4 Análise de qualidade e pré-processamento das reads	30
3.5 Montagem dos transcritos	31
3.6 Filtragem das montagens	32
3.7 Anotação	33
3.8 Avaliação de qualidade e predição de ortologia	34
3.9 Cálculo de abundância dos transcritos	36
3.10 Análises de expressão diferencial	37
<b>4. RESULTADOS</b>	<b>40</b>
4.1 Montagem <i>de novo</i> do transcriptoma de <i>P. platycephala</i> e <i>S. pulcherrimum</i>	40
4.1.1 – Qualidade dos RNAs e das bibliotecas de cDNA.	40
4.1.2 – Sequenciamento e avaliação de qualidade das reads geradas.	43
4.1.3 – Montagem dos transcriptomas de referência.	45
4.1.4 – Verificação de qualidade e predição de ortólogos.	48
4.2 Plasticidade da expressão gênica nos substratos naturais	50
4.2.1 – Sobre o número de réplicas biológicas.	50
4.2.2 – Relação entre as amostras	51
4.2.3 – Análises de expressão diferencial	52

4.3	Respostas transcriptômicas ao crescimento em substratos de estéril de mina-----	57
4.3.1	– Relação entre as amostras-----	57
4.3.2	– Análises de expressão diferencial-----	59
5.	DISCUSSÃO-----	69
6.	CONCLUSÕES-----	80
7.	REFERÊNCIAS-----	82
8.	CONSIDERAÇÕES FINAIS-----	94
8.1	Premiação-----	95
8.2	Premiação suplementar-----	95
8.3	Produção científica suplementar no período de doutoramento-----	95
9.	APÊNDICES-----	97
9.1	APÊNDICE 1: Lista completa de Termos GO enriquecidos nos GDE de <i>P. platycephala</i> cultivada em substrato de canga.-----	97
9.2	APÊNDICE 2: Lista completa de Termos GO enriquecidos nos GDE de <i>S. pulcherrimum</i> cultivada em substrato de canga.-----	100
9.3	APÊNDICE 3: Lista completa de Termos GO enriquecidos nos GDE de <i>P. platycephala</i> cultivada em estéril amarelo.-----	104
9.4	APÊNDICE 4: Lista completa de Termos GO enriquecidos nos GDE de <i>S. pulcherrimum</i> cultivada em estéril vermelho.-----	105
9.5	APÊNDICE 5: Lista completa de Termos GO enriquecidos nos GDE de <i>S. pulcherrimum</i> cultivada em estéril amarelo.-----	107
9.6	APÊNDICE 6: Artigo publicado na revista <i>BMC Genomics</i> contendo as partes 1 e 2 do presente trabalho.-----	109

## 1. Introdução

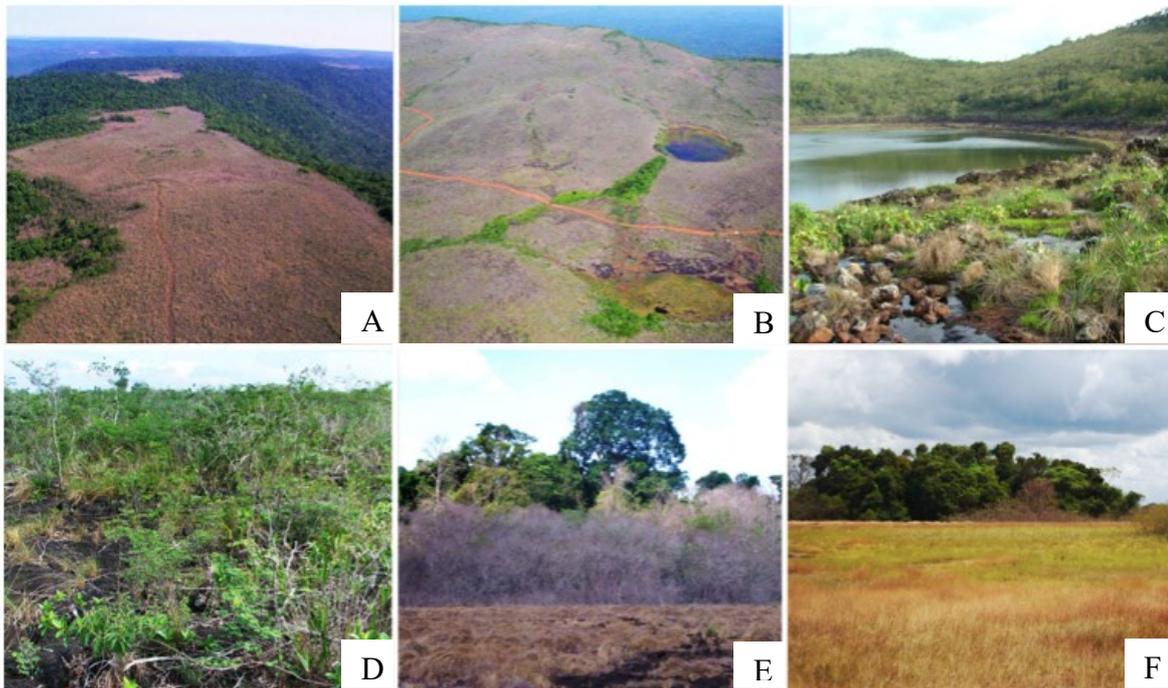
### 1.1 A canga

As áreas montanhosas são centros de endemismo e diversidade reconhecidos, especialmente devido ao seu isolamento e variação de altitude. No Brasil, afloramentos ferruginosos, conhecidos como canga, partilham características com outros afloramentos rochosos, como isolamento, alta exposição aos raios ultravioleta, grandes variações térmicas diárias, ventos constantes, alta evapotranspiração, baixa retenção de água e solos superaquecidos e impermeáveis (JACOBI, 2007). Além disso, as cangas contêm altos teores de óxidos de ferro, atingindo níveis acima de 80% do total de óxidos, e estão no centro das atenções de várias empresas de mineração (SILVEIRA, 2016).

As cangas são formações superficiais compostas por minerais derivados de formações ferríferas bandadas (SIMMONS, 1963). Podem ser originadas a partir de transporte, deposição, processos químicos e mecânicos, bem como do enriquecimento de ferro (Fe), alumínio (Al) e outros elementos residuais, conforme observado em regiões tropicais do Brasil e do mundo (DORR, 1969; SCHAEFER et al., 2015). Alguns afloramentos de canga podem ter cerca de 20 a 30 metros de espessura, com registros de origem de cerca de 50 milhões de anos, representando assim uma das paisagens mais antigas do Brasil (DORR, 1969; SALGADO E CARMO 2015).

Por sua distribuição restrita, os afloramentos ferruginosos são ecossistemas singulares, com presença de espécies ameaçadas de extinção ou endêmicas, ocorrendo sobre jazidas de minério de ferro economicamente importantes. No Brasil, ocorrem majoritariamente no quadrilátero ferrífero (MG), Serra dos Carajás (PA), Caetité (BA) e Morraria de Urucum (MS) (CARMO et al., 2012). Nessas regiões, os solos de ocorrência possuem muitas peculiaridades que revelam uma adaptação extraordinária das plantas a um cenário edáfico extremamente restritivo, tanto físico, quanto quimicamente (SCHAEFER et al., 2015).

Inserida no domínio fitogeográfico da Amazônia, a região de Carajás abriga vegetação campestre numa matriz dominada por formações florestais que variam desde as florestas ombrófilas sempre verdes até as estacionais, com diferentes graus de deciduidade (SCHAEFER et al., 2015). Sobre os platôs cobertos por afloramentos ferruginosos na região, encontra-se um rico mosaico de fitofisionomias (Figura 1), entre os quais se destacam a superfície rochosa, fendas e depressões, capões, pequenos lagos temporários e cavernas (JACOBI et al., 2007).



**Figura 1:** Fitofisionomias associados às cangas da Serra dos Carajás – A e B - fotografias aéreas de afloramentos ferruginosos, Serra Sul; C - vegetação rupestre em borda de lagoa, na Serra Sul; D - vegetação rupestre arbustiva sobre canga; E - vegetação rupestre e transição para capão florestal, estação seca; F - campos limpos e capão florestal, estação chuvosa. (Fotos: A e B - Leonardo Vianna; C, D e E - Pedro Viana; F - Nara Mota). Figura adaptada de Viana et al. (2016).

A vegetação associada à canga exibe uma série de modificações fisiológicas para adaptação a ambientes extremos. Devido ao pouco volume de solo, as plantas são altamente dependentes da ciclagem de nutrientes e desenvolveram mecanismos de tolerância a estresse hídrico, térmico e salino (JACOBI et al., 2007). Além disso, a associação com microrganismos benéficos é essencial para estabelecimento e sucesso das plantas neste tipo de ambiente (CICATELLI et al., 2010; VAN DER BIJ et al., 2018). As principais características das plantas adaptadas ao ambiente da canga envolvem folhas grossas, cerosas, coriáceas ou com tricomas, estômatos protegidos e controle da atividade estomática, idioblastos contendo compostos fenólicos e cristais, entre outras características. (GIULIETTI et al. 1987; JACOBI et al., 2007; SKIRYCZ et al., 2014). A comunidade de plantas de canga também exibe adaptações para viver sobre um substrato rico em metais, como evitar a absorção do solo, exclusão ou acúmulo nas folhas (VIEHWEGER, 2014). O acúmulo de metal foi registrado em várias espécies nas cangas de Carajás (SILVA, 1992). Até onde sabemos, não há estudos abordando os mecanismos moleculares relacionados ao estabelecimento das plantas no ambiente da canga.

## 1.2 Mineração e a reabilitação das áreas degradadas

A singularidade das cangas, alta diversidade, falta de estudos e pressão econômica sobre o ecossistema são desafios para sua conservação. A perda de área de canga está associada à exploração de recursos naturais, uma vez que abrigam as maiores reservas globais de minério de ferro, além de manganês e bauxita (alumínio). O desenvolvimento dos projetos de mineração na região de Carajás, por exemplo, causou uma redução de 19,6% na área de canga original (SOUZA-FILHO et al., 2019). Depósitos de minério de ferro encontram-se associados aos principais geossistemas ferruginosos no Brasil e eles representam cerca de 16,8% das reservas lavráveis globais. Essas reservas estão concentradas em apenas três Estados: Minas Gerais (81,6% das reservas, teor do minério em torno de 43,7% de Fe, representando 63,9% da produção de minério no Brasil), Mato Grosso do Sul (2,7% das reservas, teor do minério em torno de 62,6%, representando 0,8% da produção de minério no Brasil) e Pará (10,6% das reservas, teor do minério em torno de 65,0%, representando 35,2% da produção de minério no Brasil) (SOUZA & CARMO, 2015; DNPM, 2017). Essa peculiaridade geoeconômica favorece a concentração de megaestruturas para extração de minério (JACOBI et al., 2011; SONTER et al., 2014).

O impacto direto da mineração sobre as superfícies terrestres é grande, com a destruição dos ecossistemas naturais pela remoção do solo, plantas e animais ou pela supressão devido às pilhas de resíduo (COOKE & JOHNSON, 2002). Na restauração de locais onde o solo superficial foi perdido, os principais desafios ecológicos são com relação à interação entre as plantas e os substratos. Assim, a seleção e propagação de espécies promissoras para fins de revegetação requerem informações detalhadas sobre a ecologia das espécies (GIANNINI et al., 2016).

Em comparação com os solos não impactados, os substratos de mineração, ou os resíduos produzidos a partir do processamento dos minerais, podem apresentar desafios extremos à colonização por plantas. Quase todos os substratos de mina têm níveis muito baixos de macronutrientes (especialmente nitrogênio (N), fósforo (P) e potássio (K)) e baixo pH. Toxicidade, especialmente de alumínio, zinco e outros metais em resíduos ácidos, pode ser um problema significativo para o crescimento das plantas (COOKE & JOHNSON, 2002). Dessa forma, a revegetação e a reabilitação das áreas de mina ainda são desafiadoras. Os principais desafios incluem a seleção, propagação e restabelecimento de espécies promissoras; o controle de espécies invasoras e o monitoramento do sucesso da reabilitação (SKIRYCZ et al., 2014; MCDONALD et al., 2016; GASTAUER et al., 2018). Apesar dos desafios, a revegetação na

área de Carajás é uma realidade e o conhecimento científico e prático adquiridos têm papel central no seu aprimoramento.

O rápido crescimento e fechamento da vegetação são necessários para o controle da erosão e a estabilização do local. Assim, espécies vegetais de crescimento rápido são necessárias (GASTAUER et al., 2018). O uso de espécies nativas durante os projetos de restauração é desejável e a propagação de espécies endêmicas e ameaçadas contribui para o cumprimento dos objetivos de conservação. Giannini e colaboradores (2016) elaboraram um plano de recuperação para as áreas degradadas das cangas da Serra de Carajás. Foram selecionadas 10 características de plantas nativas para guiar programas de restauração: Uso potencial como plantas-berçário (plantas facilitadoras); capacidade reprodutiva por semente; anemocoria; presumidamente endêmicas, raras ou ameaçadas; ocorrência em mais de um afloramento de Canga; ocorrência na Canga e na floresta associada; polinização por animais; zoocoria; uso potencial pela comunidade local; espécie-bandeira para o ecossistema da Canga. Com base nessas características, os autores propuseram uma lista de 53 espécies classificadas para restauração de Canga, sendo Euphorbiaceae, Leguminosae (Fabaceae) e Orchidaceae as famílias com maior número de representantes.

No Brasil, as empresas de mineração devem apresentar ao órgão ambiental um Plano de Recuperação de Áreas Degradadas (PRAD), documento que preconiza a adoção de procedimentos para estabelecer ou restabelecer a cobertura vegetal nas áreas degradadas. A prática de revegetação é necessária para compensar as perdas líquidas de recursos bióticos das operações de mineração (ALMEIDA & SANCHEZ, 2005). Os processos de licenciamento exigem informações básicas sobre a biota e os serviços ambientais associados. As atividades de mineração devem ser conduzidas no interesse de controlar sua interferência no meio ambiente. O atendimento às demandas legais trouxe a necessidade da realização de pesquisas que contribuam com os desafios das empresas e com o conhecimento científico sobre a flora na canga de Carajás (VIANA et al., 2016). Existem grandes lacunas no conhecimento sobre as comunidades vegetais e de serviços ecológicos com potencial de uso em áreas afetadas direta ou indiretamente pela extração de minério de ferro. O conhecimento das estratégias de crescimento das plantas e outros fatores que podem afetar a dinâmica de recuperação ou reabilitação da vegetação de canga ainda é muito limitado (GASTAUER et al., 2018).

### **1.3 Plasticidade da expressão gênica e a seleção de espécies para recuperação**

Uma revegetação bem-sucedida requer que os organismos sejam capazes de se estabelecer e persistir no novo habitat. Quando os indivíduos são introduzidos, eles enfrentam novas pressões seletivas impostas pelo novo ambiente. Assim, a sobrevivência de indivíduos reintroduzidos depende no longo prazo da adaptação dos organismos direcionada pela diversidade genética. Poucos estudos incluem medidas de diversidade genética na avaliação de conservação das espécies (RIVERS et al., 2014; WILLOUGHBY et al., 2015; BABIYCHUK et al., 2017; LANES et al., 2018; CARVALHO et al., 2019), embora a diversidade genética determine a capacidade de adaptação às mudanças ambientais e, portanto, sustente a persistência das espécies. A permanência e sobrevivência dos organismos em ambientes em constante mudança depende também da sua capacidade de exibir fenótipos alternativos em resposta a diferentes condições (HE et al., 2016). Assim, no curto prazo, os indivíduos reintroduzidos podem sobreviver por meio da aclimação ambiental via plasticidade fenotípica (SCHLICHTING & SMITH. 2002; GHALAMBOR et al. 2007). A avaliação da plasticidade da expressão gênica na escolha das espécies ou populações pode aumentar a taxa de sucesso de recuperação. A variação na expressão gênica é um processo fisiológico e importante mecanismo adaptativo que induz plasticidade fenotípica e permite a resposta dos organismos a novos habitats (SCHULTE, 2004). A plasticidade fenotípica induzida pelo ambiente desempenha papéis importantes no desenvolvimento, função e adaptação dos organismos. Idealmente, os pesquisadores devem avaliar as populações-fonte potenciais para combinar as características do habitat, a arquitetura genética associada e variação fenotípica para escolher a melhor fonte possível para reintrodução (SORK et al. 2013).

Como a variação na expressão gênica está diretamente ligada à variação fenotípica, a avaliação da expressão gênica diferencial é uma ferramenta poderosa para a previsão de fenótipos (OELLRICH et al. 2014). Diferentes espécies ou mesmo populações geralmente respondem a estressores ambientais de maneiras diferentes no nível da transcrição gênica. A aplicação desse conhecimento para a seleção de espécies ou populações fonte para reintroduções é praticamente inexistente (HE et al., 2016).

A identificação de recursos genéticos naturais e a caracterização da plasticidade da expressão gênica em plantas de canga podem aumentar nosso conhecimento sobre os genes que respondem a diversos estresses abióticos, como metais pesados, deficiência de nutrientes, seca e calor. Esse conhecimento pode ser aplicado ao uso de espécies nativas e/ou endêmicas na reabilitação de áreas (GASTAUER et al., 2020) e fornecer indícios de respostas adaptativas relevantes para espécies de importância agrícola (NTIAMOA-BAIDU, 1997).

Uma forma eficiente de analisar os efeitos de condições de estresse em plantas se dá pela avaliação do seu transcriptoma. Estudos sobre a composição do transcriptoma em diferentes células, tecidos ou órgãos, em diferentes condições ambientais e espaço-temporais podem contribuir para a identificação de genes importantes para o desenvolvimento e resposta das plantas ao seu ambiente (WANG et al., 2009).

#### 1.4 Plantas selecionadas

Fabaceae, também conhecida como Leguminosae, é uma das maiores famílias de plantas encontradas nas cangas da Serra dos Carajás, com quase 80 espécies (MATTOS et al., 2018). A família compreende culturas de importância socioeconômica, sendo a segunda família de plantas mais cultivada (KOENEN et al., 2020) e inclui aproximadamente 19.500 espécies divididas em seis subfamílias recentemente descritas: Cercidoioideae, Detarioideae, Dialioideae, Duparquetioideae, Papilionoideae, e Caesalpinioideae, que inclui a antiga subfamília Mimosoideae (LPWG, 2017). A característica ecológica mais proeminente da família é a possibilidade de algumas espécies de leguminosas fixarem eficientemente o nitrogênio atmosférico em simbiose com bactérias do solo (TEDERSOO et al., 2018). Dessa forma, as Fabaceae obtêm um maior êxito na colonização de solos pobres nestes nutrientes, tornando-se muito promissoras na recuperação de áreas degradadas (FARIA et al. 2011). No contexto de Carajás, a família vem sendo estudada quanto à sua capacidade de nodulação e potencial de uso em recuperação de áreas alteradas pela mineração (FARIA et al. 2011; SILVA et al., 2018; MATTOS et. al., 2018).

*Parkia platycephala* Benth. e *Stryphnodendron pulcherrimum* (Willd.) Hochr. são duas espécies de Caesalpinioideae tropicais com faixas de distribuição compreendendo diferentes regiões, biomas e habitats, com ocorrências naturais em ambientes de canga e nas florestas circundantes. *P. platycephala* é uma espécie endêmica do território brasileiro ocorrendo nos domínios da Amazônia, Caatinga e Cerrado (FORZZA et al., 2010), enquanto *S. pulcherrimum* é mais amplamente distribuída no continente sul-americano, cobrindo grandes domínios fitogeográficos da Amazônia, Caatinga e Mata Atlântica (FORZZA et al., 2010). Ambas as espécies são adequadas para programas de reabilitação de minas (GIANNINI et al., 2016, SILVA et al., 2018, RAMOS et al., 2019, GASTAUER et al., 2020), exibindo taxas de crescimento rápido e tolerância à seca (SILVA et al., 2018, LORENZI, 2002), além de *S.*

*pulcherrimum* ser sabidamente uma espécie fixadora de nitrogênio associada a rizóbios (FARIA et al. 2011).

Devido a alta heterogeneidade de substratos em ambientes de mineração, é necessário determinar o desempenho das espécies vegetais e suas interações nos mais diversos ambientes a serem reabilitados (BARR et al., 2017). Para melhorar a reabilitação ambiental após a extração de minério de ferro na Província Mineral de Carajás, Silva e colaboradores (2018) delinearão as influências de quatro diferentes substratos sobre o desenvolvimento inicial de *P. platycephala* e *S. pulcherrimum*: dois substratos naturais (onde há ocorrência das espécies - solo de canga, solo de uma área de mata adjacente à canga) e dois substratos originados das pilhas de estéril da Vale Serra Norte (estéril vermelho e estéril amarelo). Os autores observaram que as plantas cultivadas nos substratos naturais sem fertilização não apresentaram diferença na taxa de crescimento inicial, mas as plantas cultivadas no substrato de canga apresentaram maior investimento relativo no desenvolvimento das raízes e maiores concentrações foliares de manganês (Mn) e ferro (Fe), indicando o potencial das espécies como hiperacumuladoras de metal (SILVA et al., 2018). Nos substratos de estéril, houve menor acúmulo de biomassa de *S. pulcherrimum* em estéril amarelo. Entre as duas espécies, apenas *S. pulcherrimum* exibiu interação simbiótica com bactérias fixadoras de nitrogênio. Essa interação foi observada em todos os substratos, com exceção de estéril amarelo, e teve maior incidência em substrato de mata.

Determinar como as plantas se adaptaram a ambientes extremos pode fornecer informações úteis sobre a forma como respondem ao estresse. O estudo da plasticidade da expressão gênica de *P. platycephala* e *S. pulcherrimum* nos dará informações relevantes para a compreensão de mecanismos adaptativos destas espécies à canga, além de gerar informações genéticas inéditas acerca das espécies. Os resultados serão também importantes para o monitoramento da resposta das plantas, seleção de variedades que melhor se adaptam a áreas sob processo de recuperação e de melhor preparo e escolha do substrato.

Nesse trabalho, inicialmente avaliamos a plasticidade da expressão gênica nos substratos em que as espécies são adaptadas (canga e mata – chamados aqui de substratos naturais) e identificamos genes ortólogos comuns às espécies que respondem de maneira similar aos substratos para avaliar se esses genes têm relação com o sucesso dessas espécies no ambiente da canga.

Na segunda parte, avaliamos o desempenho das espécies nos substratos de áreas degradadas, além de avaliar se conseguimos prever o sucesso das espécies na ocupação das áreas. Como as duas espécies apresentaram desempenhos diferentes nos substratos de estéril,

também avaliamos os ortólogos diferencialmente expressos entre as espécies com interação com os substratos de mina. Esses genes podem estar relacionados às diferentes respostas adaptativas observadas nas espécies.

## 2. Objetivos

### 2.1 Objetivo geral

- Determinar os mecanismos moleculares envolvidos na adaptação de *Parkia platycephala* e *Stryphnodendron pulcherrimum* a diferentes habitats associados à atividade mineradora.

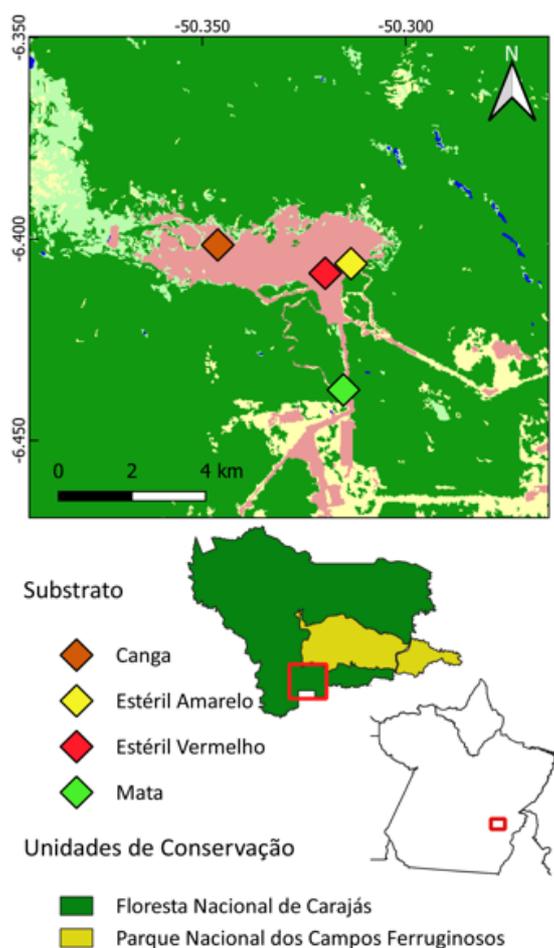
### 2.2 Objetivos específicos

- Montar transcriptomas de referência para *P. platycephala* e *S. pulcherrimum*;
- Determinar a plasticidade de expressão gênica das espécies nos substratos naturais;
- Identificar genes ortólogos entre as espécies para facilitar a comparação dos perfis de expressão;
- Identificar processos biológicos e funções moleculares alterados frente aos diferentes tipos de solo;
- Avaliar a resposta das espécies frente as condições adversas dos substratos de estéril.

### 3. Materiais e Métodos

#### 3.1 Material biológico e desenho experimental

O material biológico foi gentilmente fornecido pelo grupo de Recuperação de Áreas Degradadas (RAD) do Instituto Tecnológico Vale (ITV). Em resumo, *P. platycephala* e *S. pulcherrimum* foram cultivadas em quatro substratos coletados na Província Mineral de Carajás (CMP), Pará, Brasil. Uma área de canga, uma área de mata adjacente à canga e dois substratos originados das pilhas de estéril da Vale Serra Norte (estéril vermelho e estéril amarelo) foram escolhidos para a coleta de substrato (SILVA et al., 2018). Os pontos de coleta e os atributos físico-químicos dos substratos podem ser observados na Figura 2 e na Tabela 1, respectivamente.



**Figura 2:** Mapa evidenciando o local de coleta dos diferentes substratos utilizados no experimento.

**Tabela 1:** Atributos químicos e físicos dos substratos coletados. Tabela adaptada de Silva e colaboradores (2018).

Atributos	Solo de Mata	Solo de Canga	Estéril Vermelho	Estéril Amarelo
pH em H <sub>2</sub> O	5,0	4,9	4,7	4,7
P (mg dm <sup>-3</sup> )	3,1	18	11,5	2,9
K (mg dm <sup>-3</sup> )	8,3	11,5	8,4	11,7
S (mg dm <sup>-3</sup> )	16,3	28,1	76,9	154
Ca (cmol <sub>c</sub> dm <sup>-3</sup> )	1,6	0,1	0,2	0,1
Mg (cmol <sub>c</sub> dm <sup>-3</sup> )	1,2	0,1	0,1	< 0,1
Al (cmol <sub>c</sub> dm <sup>-3</sup> )	< 0,1	< 0,1	< 0,1	< 0,1
Zn (mg dm <sup>-3</sup> )	0,2	1,9	0,1	4,0
Fe (mg dm <sup>-3</sup> )	111	108,3	129,4	99
Mn (mg dm <sup>-3</sup> )	1,2	13,1	5,2	1,1
Cu (mg dm <sup>-3</sup> )	0,8	0,2	0,6	8,6
B (mg dm <sup>-3</sup> )	0,1	< 0,1	< 0,1	< 0,1
Argila (%)	37,0	15,0	39,0	69,0
Silte (%)	24,4	13,8	23,4	15,9
Areia (%)	38,6	71,2	37,6	15,1
Classificação	Franco-Argiloso	Franco-Arenoso	Franco-Argiloso	Argila Pesada

Ambas as espécies foram cultivadas em casa de vegetação em bandejas de 35 cm de comprimento por 24 cm de largura e 18 cm de altura, contendo 12 L de um dos substratos. Três mudas de plantas de cada espécie em cada bandeja foram consideradas uma única réplica. Após 45 dias, folhas dos indivíduos de *P. platycephala* e *S. pulcherrimum* cultivados em duas bandejas de cada substrato foram coletadas, imediatamente congeladas em nitrogênio líquido e armazenadas em freezer -80 °C para posterior extração do RNA. O grupo de RAD foi responsável pela coleta e análise dos substratos, cultivo e análise de crescimento das espécies e coleta do material biológico. Uma descrição detalhada do local de coleta dos substratos e do cultivo das plantas está disponível em Silva e colaboradores (2018).

A seção de resultados do presente trabalho é dividida em três partes: 1 – Montagem dos transcriptomas foliares completos de *P. platycephala* e *S. pulcherrimum* para uso como referência; 2 – Plasticidade da expressão gênica das plantas cultivadas em substratos onde ocorrem naturalmente; 3 – Análise de expressão diferencial das plantas submetidas aos substratos dos resíduos de mineração. Dessa forma, a extração de RNA, sequenciamento e montagem dos transcriptomas foram realizados com amostras de ambas as espécies cultivadas nos quatro substratos. As análises da parte 2 foram realizadas apenas com as amostras de plantas cultivadas nos substratos de canga e mata, utilizando as amostras de mata como controle, e as análises da parte 3 foram realizadas com as amostras de plantas cultivadas nos substratos de canga e nos substratos de estéril de mina, utilizando as amostras de canga como controle.

### 3.2 Extração de RNA

O RNA total foi extraído utilizando-se o kit RNeasy Plant Mini kit (QIAGEN) seguindo as instruções do fabricante com modificações. As folhas coletadas de plantas crescendo na mesma bandeja foram agrupadas, maceradas em nitrogênio líquido e consideradas uma réplica de cada espécie. Ao macerado foi acrescentado 600  $\mu$ L do tampão de extração RLT. Após homogeneização, foi adicionado 0,5 volume de etanol 100%. A mistura foi transferida para a coluna RNeasy Mini spin e centrifugada por 30 segundos a 10.000 rpm. Após descarte do filtrado, o tampão RW1 foi adicionado à coluna e novamente centrifugado (10.000 rpm por 30s). O filtrado foi novamente descartado e à coluna foi adicionado o tampão RPE seguindo de centrifugação (10.000 rpm por 30s). Esse passo foi repetido mais uma vez, seguido por centrifugação a 10.000 rpm por 2 min. A eluição do RNA foi feita com 30  $\mu$ L de água RNase-free fornecida pelo kit.

O RNA total extraído foi quantificado no Qubit (ThermoFisher Scientific) e sua qualidade avaliada utilizando-se o 2100 BioAnalyzer (Agilent Technologies). Como resultado, foram fornecidos os valores de integridade do RNA (RNA integrity number - RIN) de cada uma das amostras, juntamente com o eletroferograma dos rRNAs (RNAs ribossomais), cuja integridade é essencial para a construção da biblioteca de RNA-seq.

### 3.3 Preparo das bibliotecas e sequenciamento

Para a montagem das bibliotecas, foram utilizadas moléculas de RNA após a depleção de rRNA, utilizando-se o kit TruSeq Stranded Total RNA with Ribo-Zero Plant Kit (Illumina), seguindo instruções do fabricante. Um total de 100 ng de RNA inicial de cada amostra foi usado nesta etapa. Todas as amostras de RNA utilizadas para o preparo das bibliotecas obtiveram RIN acima de 7,0. A cada biblioteca foi adicionado um *index* diferente, para possibilitar o sequenciamento de várias bibliotecas simultaneamente a partir de um único *pool* de bibliotecas e o reconhecimento e identificação de cada uma.

As bibliotecas foram quantificadas no Qubit (ThermoFisher Scientific) e a qualidade avaliada no 2100 BioAnalyzer (Agilent Technologies) com o kit DNA 1000. Após quantificação, as bibliotecas foram normalizadas a 4 nM considerando sua concentração e o tamanho médio dos fragmentos de DNA para que todas ficassem igualmente concentradas no pool final utilizado no sequenciamento. 5  $\mu$ l do pool das bibliotecas foi então desnaturado com

igual volume de NaOH (0,2 N) por 5 min. 5 µl de 200 mM Tris-HCL (pH 7.0) foi adicionado à reação e o pool foi diluído a 1,8 pM para posterior sequenciamento.

O sequenciamento das bibliotecas no modo *paired-end* foi realizado no equipamento Illumina NextSeq™ 500 do Instituto Tecnológico Vale - Desenvolvimento Sustentável em Belém, Pará com o kit de corrida *High Output* de 300 ciclos. Todas as *reads* brutas geradas a partir deste estudo foram depositadas no *Short Read Archive* (SRA) do NCBI sob o número de acesso PRJNA645405.

### 3.4 Análise de qualidade e pré-processamento das reads

As sequências advindas das bibliotecas de cDNAs foram demultiplexadas de forma a separá-las para cada amostra correspondente. Os indexes foram demultiplexados pelo software bcl2fastq ([https://support.illumina.com/downloads/bcl2fastq\\_conversion\\_software\\_184.html](https://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html)), que também se encarrega de remover sequências artefatos. Após demultiplexação, as *reads* foram avaliadas quanto à qualidade utilizando-se o programa FastQC (ANDREWS, 2010). Apesar de no preparo das bibliotecas haver um passo de depleção de rRNA, foi realizado o alinhamento das *reads* contra bancos de dados de rRNA do RFAM ([rfam.xfam.org](http://rfam.xfam.org)) para verificação da eficiência da remoção no preparo das bibliotecas e para retirada dos rRNAs remanescentes. Foi feito o download dos bancos de dados de rRNA 5S (RF00001), *Bacterial large subunit* (RF02541), *Bacterial small subunit* (RF00177), *Eukaryotic small subunit* (RF01960), *Eukaryotic large subunit* (RF02543) e 5.8S (RF00002). As sequências dos bancos foram concatenadas em um único arquivo e indexadas com bowtie-build. O programa Bowtie v. 1.2.2 (LANGMEAD et al., 2017) foi utilizado para o alinhamento das *reads* contra os bancos de dados com o parâmetro --un (arquivo com todas as *reads* que não mapearam à referência). As *reads* que não mapearam às sequências no banco de dados indexado foram utilizadas para as análises posteriores. Para remoção de adaptadores, bases indefinidas (Ns) e sequências de baixa qualidade foi utilizado o Trimmomatic v0.38 (BOLGER et al., 2014) com os seguintes parâmetros: ILLUMINACLIP:TruSeq3-PE-2.fa:2:20:10 (remove adaptadores Illumina) LEADING:3 (remove bases de qualidade inferior a 3 a partir do início da sequência até encontrar uma base com valor maior do que esse) TRAILING:3 (remove bases de qualidade inferior a 3 a partir do final da sequência até encontrar uma base com valor maior do que esse) SLIDINGWINDOW:5:20 (percorre a *read* com uma janela deslizante de 5 bases e elimina

janelas cuja qualidade média seja inferior a 20) MINLEN:35 (elimina *reads* cujo tamanho seja inferior a 35 após os passos realizados acima).

### 3.5 Montagem dos transcritos

As *reads* de alta qualidade de *P. platycephala* e *S. pulcherrimum* foram montadas *de novo* utilizando quatro montadores com diferentes parâmetros e tamanhos de k-mer, na tentativa de maximizar a diversidade e produzir o conjunto de transcritos de plantas mais preciso e completo possível. Os pares de *reads* resultantes da trimagem pelo Trimmomatic foram utilizadas por todos os montadores, exceto para as montagens geradas com SOAPDenovo-Trans, que também incluiu as *reads single-end* de alta qualidade resultantes. Os montadores utilizados foram: Trinity v.2.8.3 (GRABHERR *et al.*, 2011), com os parâmetros --seqType fq (formato dos arquivos de *input* - fastq); --max\_memory 260G (sugestão de máximo de memória a ser utilizado pelo Trinity quando o uso de memória pode ser limitado); --left (*reads* 1 separadas por vírgula); --right (*reads* 2 separadas por vírgula); --SS\_lib\_type RF (tipo de biblioteca *strand-specific* – RF = *read* 1 sequenciada no sentido antisenso e *read* 2 no sentido senso); --bflyCalculateCPU (módulo *Butterfly* calcula o número de CPUs a serem utilizados com base em 80% do máximo de memória sugerido); --group\_pairs\_distance 100 (comprimento esperado entre os pares de *reads*); --normalize\_by\_read\_set (normaliza os pares de *reads* separadamente, depois realiza uma normalização final que combina as *reads* normalizadas individualmente); maSPAdes v. 3.12.0 (BANKEVICH *et al.*, 2012), com os parâmetros --ss-rf (tipo de biblioteca *strand-specific* – rf = *read* 1 sequenciada no sentido antisenso e *read* 2 no sentido senso); Velvet v. 1.2.10/Oases v. 0.2.09 (SCHULZ *et al.*, 2012), com os parâmetros -shortPaired (tipo de read – curta e *paired-end*) -fastq (formato do arquivo) -strand\_specific (tipo de biblioteca) para velveth e -ins\_length 330 (tamanho do fragmento sequenciado) para oases; e SOAPDenovoTrans v. 1.03 (XIE *et al.*, 2013), com os parâmetros max\_rd\_len = 150 (tamanho máximo das *reads*), avg\_ins = 300 (tamanho médio do fragmento sequenciado), reverse\_seq = 0 (reverso complemento desnecessário), asm\_flags = 3 (usar as *reads* para a montagem de *contigs* e *scaffolds*), map\_len = 32 (comprimento mínimo de alinhamento entre uma *read* e um *contig* – Para *paired-end* = 32) no arquivo de configuração. Kmergenie v. 1.7039 (CHIKHI & MEDVEDEV, 2013) foi usado para calcular o provável melhor tamanho de k-mer para cada montagem a partir da união de todas as *reads forward* das bibliotecas, todas as *reads reverse* e a união de todas as *reads (forward e reverse)*. Trinity foi

utilizado com k-mer 25, rnaSPAdes calcula o melhor tamanho de k-mer per si e por isso nenhum valor de k-mer foi especificado para o programa. Velvet/Oases e SOAPDenovo-Trans foram utilizados com múltiplos k-mers, variando de 10 em 10 a partir de 21 até 81 além dos k-mers resultantes do Kmergenie (31, 51 e 57 para *P. platycephala*; 27, 31 e 39 para *S. pulcherrimum*).

As montagens resultantes foram combinadas em uma montagem única para cada espécie e processadas para mitigar a redundância com o *pipeline* tr2aacds v.4 do EvidentialGene ([http://arthropods.eugenesis.org/EvidentialGene/about/EvidentialGene\\_trassembly\\_pipe.html](http://arthropods.eugenesis.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html); GILBERT, 2019). O *pipeline* seleciona o melhor conjunto de transcritos gerados a partir de um *input* de montagens combinadas de diferentes softwares e parâmetros baseado no potencial codificador dos *contigs*. Este pipeline usa vários programas para 1) remover redundância perfeita de sequências codificadoras (CDS) de proteínas (fastanrdb/exonerate-2.2.0); 2) remover transcritos idênticos (CD-HIT-EST); e 3) encontrar transcritos alternativos através de alinhamentos de alta identidade com exons (blastn); e classifica os transcritos em três classes: okay (transcritos primários – com CDS mais longa), okalt (transcritos alternativos, possíveis isoformas) e drop (transcritos que não passaram nos filtros internos). Posteriormente, a nova versão do programa utilizada nesse trabalho faz uma segunda etapa de análise das classes iniciais e as reclassifica em drop e okay, que contém tanto os transcritos primários quanto os alternativos. Além dos transcritos, o programa retorna as sequências das proteínas traduzidas a partir deles (proteínas preditas) e as CDS. O conjunto okay foi usado para as análises subsequentes, incluindo os transcritos primários e os transcritos alternativos.

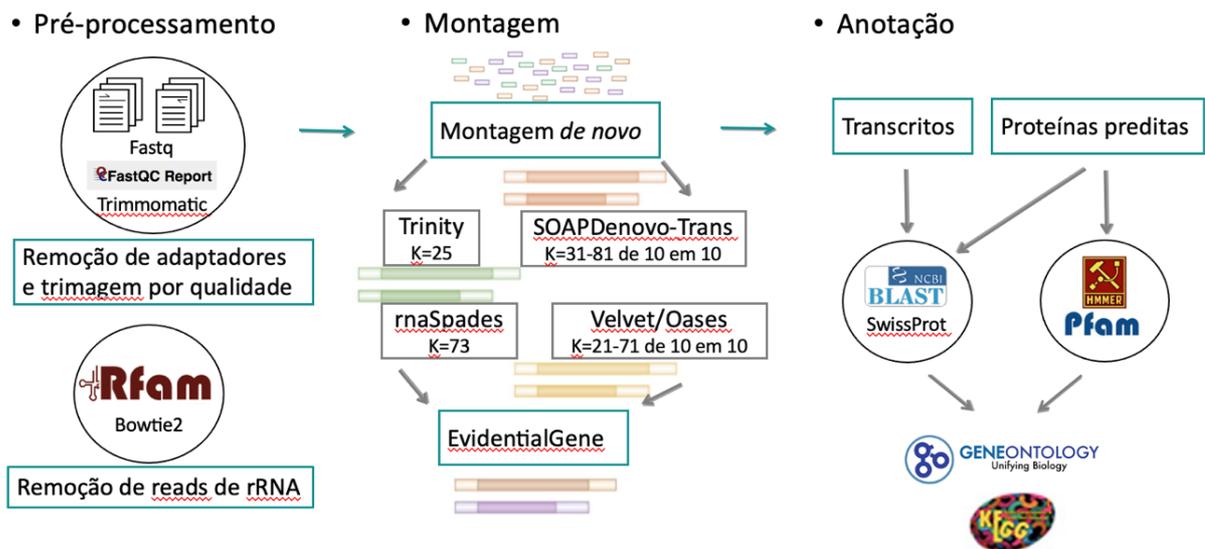
### 3.6 Filtragem das montagens

A montagem combinada a partir de múltiplos softwares e parâmetros gera transcritos mais completos e precisos, mas também adiciona muitos modelos alternativos e transcritos curtos, que codificam proteínas menores que 120 aminoácidos (aa), ao conjunto. Esses transcritos curtos são mantidos pelo pipeline do EvidentialGene por possuírem algum potencial de codificação, mas são em sua maioria loci espúrios que podem ser descartados se nenhuma evidência adicional de classificação for estabelecida (GILBERT, 2019). Dessa forma, os transcritos preliminares provenientes do EvidentialGene foram filtrados para descartar possíveis loci espúrios. Foi feito um BLAST (ALTSCHUL et al., 1990; e-value  $1 \times 10^{-6}$ ) das proteínas preditas a partir dos transcritos primários com comprimento menor que 120aa contra o banco de dados de Viridiplantae (UniProtKB / Swiss-Prot) na tentativa de estabelecer

evidências de homologia para elas. Os transcritos que codificam proteínas curtas que não obtiveram similaridade com o banco de dados de Viridiplantae foram descartados. Os transcritos curtos remanescentes e suas formas alternativas permaneceram nas montagens finais de cada espécie e foram usados para as análises de expressão diferencial.

### 3.7 Anotação

As sequências codificadoras e as proteínas preditas pelo EvidentialGene foram utilizadas para anotação funcional usando o Trinotate v.3.2.1 (BRYANT et al., 2017; disponível em <http://trinotate.github.io>). As pesquisas de homologia foram realizadas contra proteínas conhecidas usando BLASTp contra o banco de dados UniProtKB/Swiss-Prot, com um e-value de  $1e-6$ , e usando hmmer v.3.1b2 (<http://hmmer.org>) contra os domínios de proteína comuns do banco de dados Pfam. As sequências de proteínas preditas foram submetidas aos servidores SignalP 4.1 (NIELSEN, 2017) e TMHMM v.2.67 (KROGH et al., 2001), para identificação de peptídeos sinal e domínios transmembrana respectivamente. As saídas de anotação foram carregadas em um banco de dados Trinotate SQLite. A figura 3 mostra um esquema representativo das análises de bioinformática utilizadas do pré-processamento das *reads* até a anotação.



**Figura 3:** Representação da estratégia de montagem dos transcriptomas de referência. O esquema mostra a metodologia utilizada desde o pré-processamento das *reads* geradas no sequenciamento até a anotação funcional.

### 3.8 Avaliação de qualidade e predição de ortologia

Para avaliar a qualidade técnica das montagens, utilizamos o programa rnaQUAST v. 1.5.2 (BUSHMANOVA et al., 2016) e um script em perl (TrinityStats.pl) do programa Trinity. Estes scripts nos retornaram várias métricas das montagens como o número de ‘genes’, o número de transcritos e o valor de N50.

Idealmente, todas as *reads* devem ser utilizadas na montagem para obtenção dos transcritos. Assim, a medida do número de *reads* utilizadas na montagem e quantas delas são mapeadas uma única vez ao novo transcriptoma fornece pistas sobre a qualidade da montagem. Um grande número de *reads* não utilizadas indica uma montagem não ótima e um grande número de *reads* não unicamente montadas pode representar transcritos redundantes e quimeras. O programa Bowtie2 v2.2.8 (LANGMEAD & SALZBERG, 2012) foi utilizado para alinhar as *reads* utilizadas como *input* pelos montadores às montagens finais com os seguintes parâmetros: --local (alinhamento local); --nofw (modo que impede o programa de tentar mapear as *reads* 1 na fita *forward*); -x indexAssembly (prefixo da montagem indexada); -q (formato dos arquivos de entrada = fastq); -1 (caminho para as *reads* 1); -2 (caminho para as *reads* 2).

Após avaliação da qualidade técnica das montagens, o programa BUSCO v3.0.2 (*Benchmarking Universal Single-Copy Ortholog* - SIMÃO et. al., 2015) foi utilizado para avaliação da completude dos transcriptomas montados. Evolutivamente, espera-se que espécies próximas possuam um conjunto grande de genes em comum (genes ortólogos), devido ao compartilhamento de um ancestral comum recente. Caso o sequenciamento e a montagem tenham sido capazes de abranger de forma completa o transcriptoma analisado, espera-se que todos os genes ortólogos presentes em uma espécie de referência próxima possam também ser encontrados. O programa BUSCO utiliza Modelos Ocultos de Markov (HMM) para encontrar genes ortólogos de uma referência no genoma/transcriptoma analisado. Para isso, é preciso fornecer ao programa um conjunto de genes ortólogos comuns à categoria taxonômica a que o organismo sequenciado pertence. O site do programa contém conjuntos de dados de genes ortólogos para várias categorias taxonômicas incluindo plantas. Foi feito o download do banco de dados de genes ortólogos de eudicotiledôneas (eudicotyledons\_odb10) e, como os conjuntos de dados baseados no *release* 10 do OrthoDB são fornecidos na forma de PRE-RELEASE, as análises também foram feitas com o banco de dados de embriófitas (embryophyta\_odb9). Os parâmetros utilizados para execução do programa foram: -i combined\_assemblies.okay.mrna (caminho para o arquivo fasta dos transcritos a serem analisados); -l eudicotyledons\_odb10 ou

embryophyta\_odb9 (caminho para os bancos de dados de genes ortólogos de referência); -m tran (modo de análise - transcriptoma).

A qualidade da montagem do transcriptoma afeta inferências filogenéticas de forma que montagens de alta qualidade levam a reconstruções filogenéticas mais precisas, contribuindo com a consistência da filogenia já estabelecida (SPILLANE et al, 2020). Para avaliar as montagens dos transcriptomas com maior resolução e determinar as relações filogenéticas entre as espécies estudadas, examinamos a filogenia molecular da subfamília Caesalpinioideae. Para essa análise, foi feito o *download* de genomas ou transcriptomas de outras 16 espécies da subfamília a partir de várias fontes (Tabela 2), incluindo todas as sequências nucleares usadas na mais recente reconstrução filogenética da família Fabaceae (KOENEN et al., 2020) e novas sequências depositadas no banco de dados do NCBI. O programa Orthofinder v. 2.3.12 (EMMS & KELLY, 2015; 2019) foi utilizado para identificar os grupos de genes ortólogos entre as espécies a partir das sequências preditas de aminoácidos extraídas dos transcriptomas com EvidentialGene. Apenas sequências de aa provenientes dos transcritos primários foram utilizadas, uma vez que os autores recomendam o uso de um único transcrito representativo para cada gene na análise. A árvore de espécies foi inferida a partir de alinhamentos múltiplos de sequência gerados com ortogrupos de cópia única, com uma sequência presente em pelo menos 14 (77,8%) das 18 espécies analisadas. A árvore foi enraizada com *Acrocarpus fraxinifolius* como grupo externo pelo algoritmo STRIDE (EMMS & KELLY, 2017). O programa MAFFT (BENNING, 2009) foi usado para gerar os alinhamentos e o FastTree (PRICE et al., 2009) para a reconstrução da árvore.

**Tabela 2:** Espécies utilizadas na reconstrução da árvore filogenética.

Espécie	Fonte	Referência
<i>Acacia koa</i>	Genbank BioProject: PRJNA268386	Ishihara et al. 2015
<i>Acrocarpus fraxinifolius</i>	TropiTree: <a href="https://ics.hutton.ac.uk/tropiTree/">https://ics.hutton.ac.uk/tropiTree/</a>	Russel et al. 2014
<i>Albizia julibrissin</i>	Genbank BioProject: PRJNA574148	Koenen et al. 2020
<i>Chamaecrista fasciculata</i>	Dryad: <a href="http://dx.doi.org/10.5061/dryad.ff1tq">http://dx.doi.org/10.5061/dryad.ff1tq</a>	Cannon et al. 2015
<i>Desmanthus illinoensis</i>	Dryad: <a href="http://dx.doi.org/10.5061/dryad.ff1tq">http://dx.doi.org/10.5061/dryad.ff1tq</a>	Cannon et al. 2015
<i>Entada abyssinica</i>	Genbank BioProject: PRJNA574148	Koenen et al. 2020
<i>Gleditsia sinensis</i>	Genbank BioProject: PRJNA273556	Han et al. 2016
<i>Gleditsia triacanthos</i>	Dryad: <a href="http://dx.doi.org/10.5061/dryad.ff1tq">http://dx.doi.org/10.5061/dryad.ff1tq</a>	Cannon et al. 2015
<i>Gymnocladus dioicus</i>	Dryad: <a href="http://dx.doi.org/10.5061/dryad.ff1tq">http://dx.doi.org/10.5061/dryad.ff1tq</a>	Cannon et al. 2015
<i>Inga spectabilis</i>	<a href="https://doi.org/10.5061/dryad.r9c12">https://doi.org/10.5061/dryad.r9c12</a>	Nicholls et al. 2015
<i>Leucaena leucocephala</i>	Genbank BioProject: PRJNA296824	Ishihara et al. 2016
<i>Microlobius foetidus</i>	Genbank BioProject: PRJNA574148	Koenen et al. 2020
<i>Prosopis alba</i>	Genbank BioProject: PRJNA534081	Kong & Paterson
<i>Prosopis cineraria</i>	Genbank BioProject: PRJNA371524	Rai et al. 2017
<i>Senna alexandrina</i>	Genbank BioProject: PRJNA252646	Reddy et al. 2015
<i>Senna hebecarpa</i>	Dryad: <a href="http://dx.doi.org/10.5061/dryad.ff1tq">http://dx.doi.org/10.5061/dryad.ff1tq</a>	Cannon et al. 2015

### 3.9 Cálculo de abundância dos transcritos

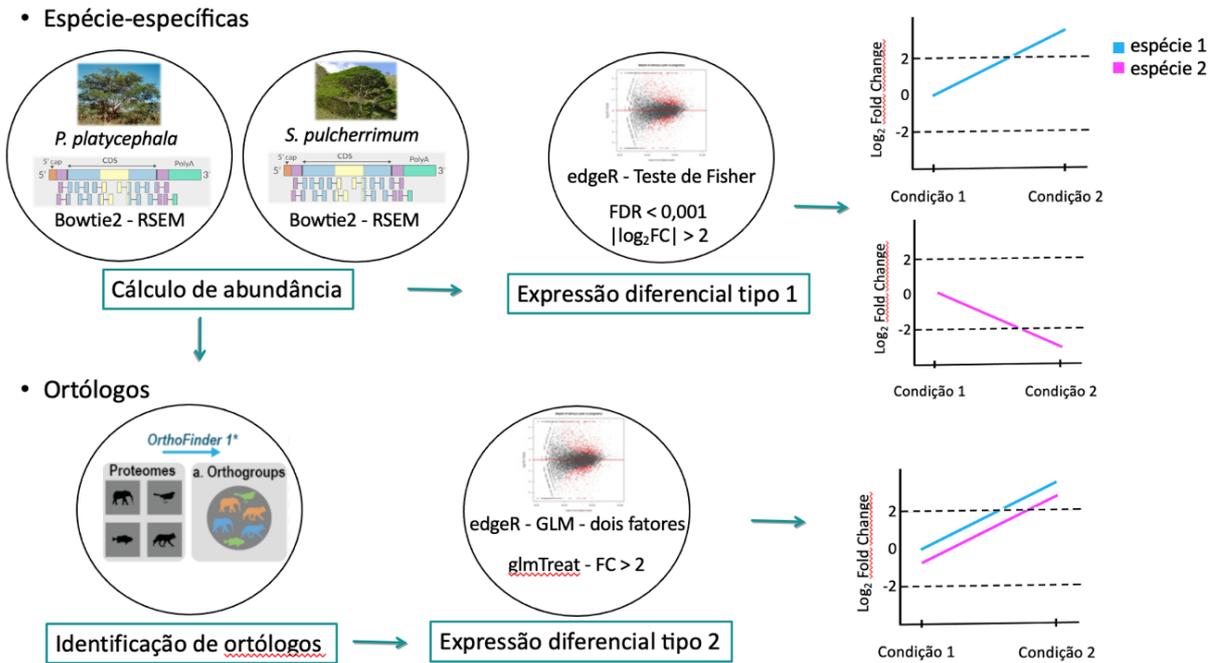
Para quantificar a abundância de transcritos de cada amostra foi utilizado o pacote RSEM v. 1.3.0 (LI & DEWEY, 2011), um método que estima a abundância de transcritos de maneira livre de genoma de referência baseado no alinhamento das *reads* ao transcriptoma montado. Um *script* do Trinity foi utilizado para executar o RSEM: `align_and_estimate_abundance.pl --transcripts Trinity.fasta` (montagem dos transcritos usada como referência); `--SS_lib_type RF` (tipo de biblioteca *strand-specific* – RF = *read* 1 sequenciada no sentido antisenso e *read* 2 no sentido senso); `--seqType fq` (formato fastq); `--samples_file` (arquivo txt com os caminhos para as *reads* e as relações de tratamento e replicatas); `--est_method RSEM` (método utilizado para estimar a abundância de transcritos); `--aln_method bowtie2` (método de alinhamento); `--gene_trans_map` (arquivo contendo os identificadores dos genes e os respectivos transcritos); `--prep_reference` (prepara a referência para o alinhamento). O RSEM foi executado a partir dos comandos acima com os seguintes parâmetros: `--paired-end` (tipo de *reads* usadas como *input* - pareadas); `--forward-prob 0` (biblioteca *strand specific* – 0 = *reads upstream*, ou seja, *reads* 1 derivadas da fita reversa); --

bam bowtie.bam (arquivo de *input* (alinhamento) no formato .bam); nome dos transcritos de referência; prefixo do nome dos arquivos de *output*.

A saída principal do RSEM consiste em dois arquivos para cada amostra, um para estimativas do nível de isoformas e outro para estimativas do nível de genes. O Trinity tem um outro *script* para facilitar a construção de uma matriz única de contagens de genes e outra de contagem de isoformas a partir dos dados de cada amostra. O *script abundance\_estimates\_to\_matrix.pl* foi utilizado, com os seguintes parâmetros: --est\_method RSEM (método utilizado para estimar as abundâncias de genes e isoformas de transcritos); --out\_prefix (prefixo do nome da matriz de *output*); --name\_sample\_by\_basedir (nomear as amostras com o nome do diretório de *input* ao invés do nome do arquivo de *input*); e os arquivos com as abundâncias de genes ou isoformas de cada amostra. As matrizes de contagem resultantes foram usadas para as análises de expressão diferencial.

### 3.10 Análises de expressão diferencial

O número de métodos e *softwares* para análise de expressão diferencial (DE) a partir de dados de RNA-Seq aumentou rapidamente. No entanto, não há consenso sobre o *pipeline* ou protocolo mais apropriado para a identificação de genes diferencialmente expressos (GDEs) a partir de dados de RNA-Seq. O pacote estatístico edgeR v. 3.8 (ROBINSON et al., 2010) foi utilizado para as análises de expressão diferencial e é baseado em modelos que seguem uma distribuição binominal negativa para estimar dados de dispersão. O edgeR funciona a partir de uma tabela de contagens de *reads*, com linhas correspondentes a genes ou transcritos e colunas para bibliotecas independentes. As contagens representam o número total de *reads* que se alinham a cada gene ou transcrito. Nas partes 2 e 3 do presente trabalho (ver seção Material biológico e desenho experimental), dois tipos de análise de expressão diferencial foram realizados, sendo o tipo 1 as análises espécie-específicas e o tipo 2 as análises de expressão diferencial dos ortólogos nas duas espécies (Figura 4). Na parte 2, as amostras cultivadas em mata foram utilizadas como controle às cultivadas em canga nos dois tipos de análise. Na parte 3, as amostras cultivadas em canga foram utilizadas como controle aos substratos de resíduo de mineração, também nos dois tipos de análise.



**Figura 4:** Representação dos diferentes tipos de análises de expressão diferencial utilizados nesse trabalho. As análises espécie-específicas (tipo 1) foram realizadas para cada espécie separadamente. As análises dos ortólogos (tipo 2) foram realizadas a partir dos genes ortólogos de cópia única encontrados entre as duas espécies com o objetivo de encontrar respostas conservadas nas espécies com relação aos desafios de substrato.

As análises espécie-específicas foram conduzidas para cada espécie separadamente afim de observar a resposta do transcriptoma como um todo aos tipos de substrato. Essas análises foram realizadas usando as contagens estimadas a partir do RSEM com o método clássico de análise do pacote edgeR (*quantile-adjusted conditional maximum likelihood* - qCML). As funções `filterByExpr` e `calcNormFactors` do edgeR foram utilizadas para filtrar genes com nível de expressão muito baixo e para normalizar as amostras pelo tamanho das bibliotecas, respectivamente. As análises foram realizadas gene a gene utilizando um teste análogo ao Teste Exato de Fisher. Os valores p resultantes foram ajustados para múltiplos testes utilizando o método de Benjamini-Hochberg (BENJAMINI & HOCHBERG, 1995), em que a taxa de erro global é controlada ajustando o *False Discovery Rate* (FDR). Os valores de corte  $FDR < 0,001$  e  $|\log_2\text{FoldChange}| > 2$  ( $|\log_2\text{FC}| > 2$ ) foram usados para classificar os transcritos como diferencialmente expressos.

As análises de expressão diferencial dos ortólogos permitem a combinação dos conjuntos de dados das duas espécies e tiveram o intuito de comparar a variação da transcrição gênica e encontrar genes comuns às duas espécies que respondem diferentemente ao tipo de substrato. As análises foram realizadas apenas com os genes ortólogos compartilhados entre as duas espécies obtidos com o OrthoFinder, conforme proposto por Moreno-Santillán e colaboradores (2019). Em resumo, os arquivos de quantificação gerados pelo RSEM para cada

espécie foram editados para substituir os IDs dos transcritos gerados pelo EvidentialGene pelo respectivo ID do ortogruppo que contém um ortólogo de cópia única em cada espécie. Os arquivos de quantificação foram concatenados em uma única matriz contendo apenas as contagens dos genes ortólogos de cópia única. A análise foi realizada com a matriz de contagem e o pacote edgeR utilizando modelos lineares generalizados (*Generalized Linear Models - GLM*) com dois fatores (espécie e substrato). Os GLMs foram implementados no edgeR usando um modelo aditivo para corrigir a diferença na expressão gênica entre as espécies e obter os genes diferencialmente expressos entre as condições (`design <- model.matrix(~Species+Substrate)`). Novamente as funções `filterByExpr` e `calcNormFactors` foram utilizadas. Para o teste de GDEs utilizamos o *quasi-likelihood F-test* (QLF), apropriado para experimentos planejados com múltiplos fatores de tratamento e com pequeno número de réplicas biológicas (CHEN et al., 2016). O método do edgeR `glmTreat` foi utilizado para classificar os ortólogos como diferencialmente expressos acima de *Fold Change 2*.

As análises de enriquecimento de ontologia gênica (GO) e de vias metabólicas foram implementadas pelo pacote GOseq do R (YOUNG et al., 2010) e pelo software KOBAS (MAO et al., 2005) respectivamente. O perfil GO do subconjunto de GDEs foi comparado ao perfil GO dos transcriptomas de referência e um valor p ajustado  $< 0,05$  foi considerado como limite de corte de significância estatística. *Medicago truncatula* foi utilizada como a espécie de background para a análise das vias metabólicas enriquecidas e um valor p ajustado  $< 0,05$  foi considerado como corte de enriquecimento. Os termos GO enriquecidos foram processados com REVIGO (SUPEK et al., 2011) para resumir as anotações principais e remover a redundância. A saída REVIGO foi então alimentada como entrada para o CircGO (KUZNETSOVA et al., 2019) para visualização. A função GOChord do pacote R GOplot (WALTER et al., 2015) também foi usada para visualizar os termos GO enriquecidos nos ortólogos DE. O pacote R ggplot2 (WICKHAM, 2016) foi usado para a visualização das vias metabólicas enriquecidas.

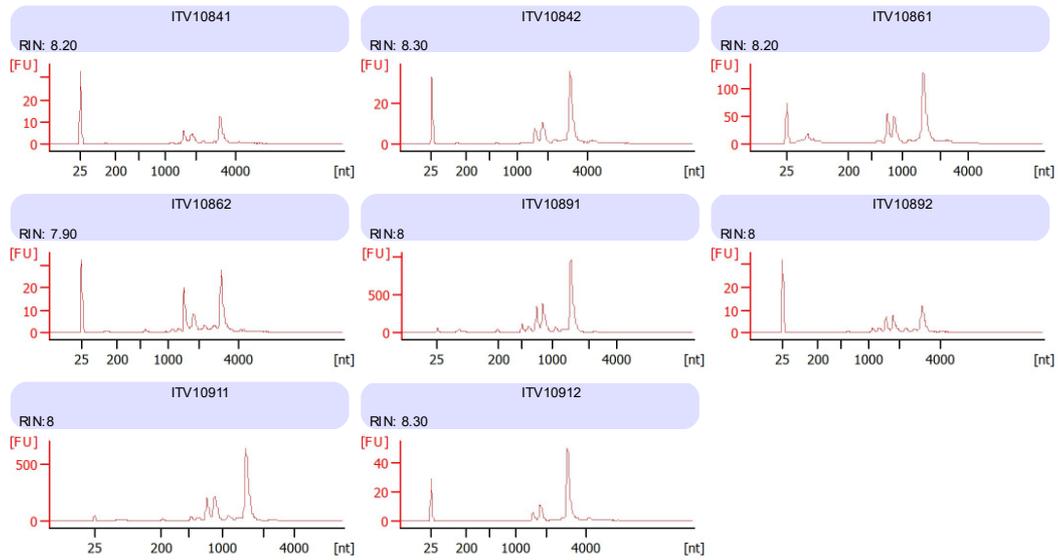
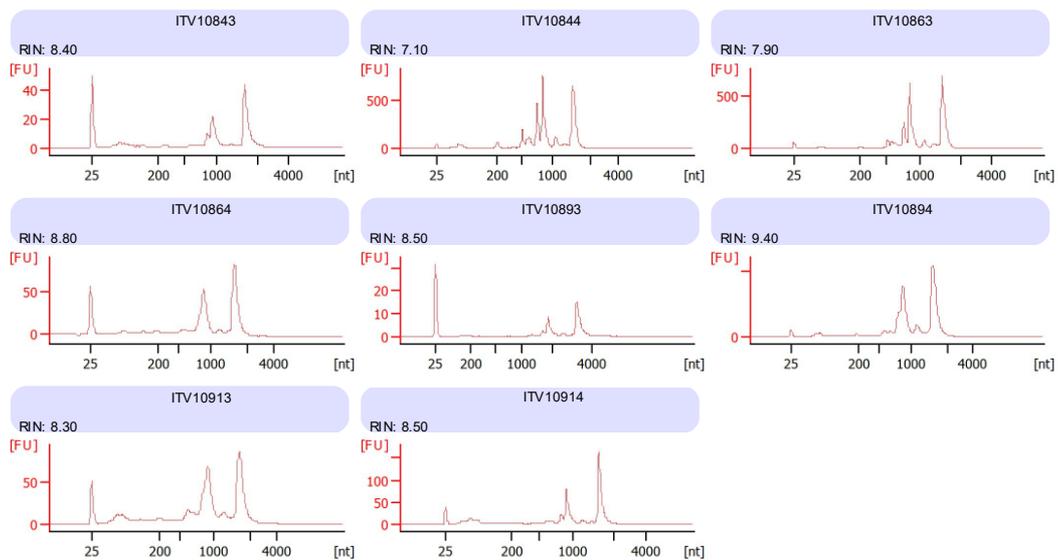
## 4. Resultados

### 4.1 Montagem *de novo* do transcriptoma de *P. platycephala* e *S. pulcherrimum*.

#### 4.1.1 – Qualidade dos RNAs e das bibliotecas de cDNA.

As leguminosas foram cultivadas na casa de vegetação da Embrapa Amazônia Oriental em Belém em quatro diferentes substratos coletados na Serra dos Carajás (canga, mata e dois resíduos de mineração). Amostras foliares de três indivíduos cultivados na mesma bandeja foram agrupadas e duas bandejas de cada substrato foram coletadas para extração de RNA.

A qualidade do RNA total extraído das plantas cultivadas em diferentes substratos foi avaliada utilizando o equipamento 2100 Bioanalyzer e os eletroferogramas resultantes podem ser observados na Figura 5. Em todas as amostras os picos de rRNA 25S e 18S foram bem resolvidos e identificados pelo software. Picos adicionais foram observados nos eletroferogramas. Esses picos adicionais são normalmente encontrados no perfil eletroforético de rRNAs de plantas uma vez que os tecidos foliares possuem rRNAs citossólicos, mitocondriais e plastidiais que variam em tamanho de 5S a 25S (BABU & GASSMANN, 2016). A concentração do RNA total de cada amostra foi medida e os resultados, junto com o RIN obtidos pelo Bioanalyzer, estão apresentados na Tabela 3.

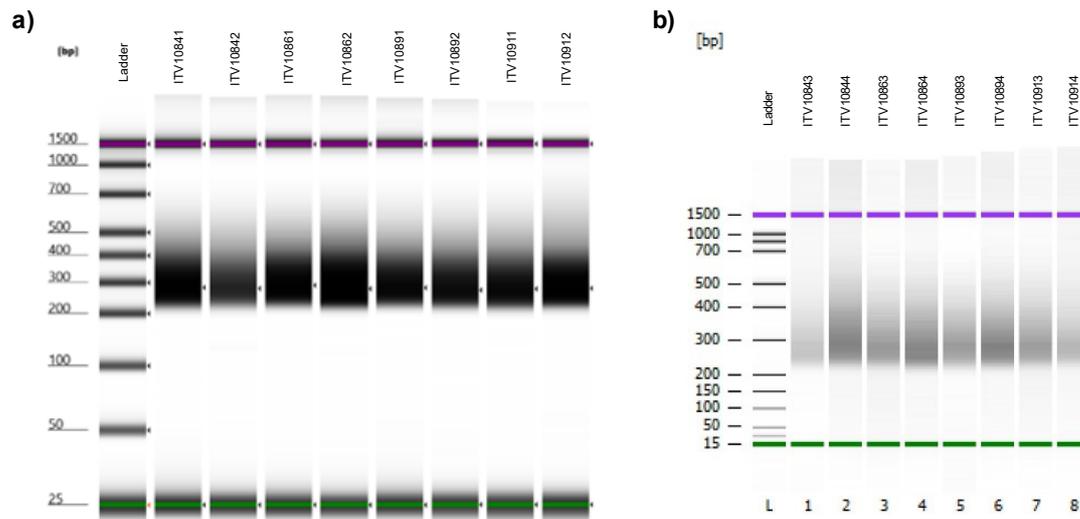
*Parkia platycephala**Stryphnodendron pulcherrimum*

**Figura 5:** Eletroferogramas das amostras de RNA de *P. platycephala* (cima) e *S. pulcherrimum* (baixo) evidenciando os picos 25S e 18S.

**Tabela 3:** Concentração e integridade dos RNAs extraídos. As amostras representam uma réplica composta de cada espécie em cada substrato.

Substrato	<i>P. platycephala</i>			<i>S. pulcherrimum</i>		
	Amostra	Concentração (ng/μL)	RIN	Amostra	Concentração (ng/μL)	RIN
Estéril amarelo	ITV10841	45,2	8,2	ITV10843	42,0	8,4
	ITV10842	67,6	8,3	ITV10844	36,2	7,1
Mata	ITV10861	13,4	8,2	ITV10863	11,1	7,9
	ITV10862	67,4	7,9	ITV10864	9,6	8,4
Canga	ITV10891	20,0	8,0	ITV10893	38,4	8,5
	ITV10892	47,8	8,0	ITV10894	14,4	9,4
Estéril vermelho	ITV10911	15,8	8,0	ITV10913	8,76	8,3
	ITV10912	60,8	8,3	ITV10914	4,8	8,5

Os RNAs totais obtidos foram utilizados para a construção de bibliotecas de fragmentos de cDNA de 330 pb de tamanho aproximadamente (Figura 6). A concentração das bibliotecas foi medida e os resultados estão apresentados na Tabela 4.



**Figura 6:** Perfil eletroforético virtual das bibliotecas de cDNA evidenciando o tamanho médio dos fragmentos obtidos. a) Perfil das bibliotecas de *P. platycephala* gerado pelo Bioanalyzer. b) Perfil das bibliotecas de *S. pulcherrimum* gerado pelo TapeStation.

**Tabela 4:** Concentração e tamanho médio de fragmentos das bibliotecas de cDNA.

<i>P. platycephala</i>			<i>S. pulcherrimum</i>		
Amostra	Tamanho médio de fragmentos	Concentração (ng/μL)	Amostra	Tamanho médio de fragmentos	Concentração (ng/μL)
ITV10841	330 pb	43,8	ITV10843	317 pb	61,6
ITV10842	335 pb	51,6	ITV10844	345 pb	55,2
ITV10861	335 pb	49,4	ITV10863	328 pb	53,0
ITV10862	332 pb	48,2	ITV10864	323 pb	54,0
ITV10891	333 pb	56,2	ITV10893	319 pb	55,6
ITV10892	329 pb	54,2	ITV10894	327 pb	59,2
ITV10911	327 pb	41,2	ITV10913	335 pb	52,6
ITV10912	336 pb	46,2	ITV10914	332 pb	56,0

#### 4.1.2 – Sequenciamento e avaliação de qualidade das *reads* geradas.

As *reads* provenientes do sequenciamento de RNA das amostras foliares foram usadas para a montagem dos transcriptomas de referência. O sequenciamento produziu entre 46 e 91 milhões de *paired-end reads* (2x 150pb) por amostra (Tabela 5).

**Tabela 5:** Resultado do sequenciamento das bibliotecas de cDNA.

<i>P. platycephala</i>				<i>S. pulcherrimum</i>		
Substrato	Amostra	Número de <i>reads</i>	Conteúdo GC	Amostra	Número de <i>reads</i>	Conteúdo GC
Estéril amarelo	ITV10841	67,271,933	44%	ITV10843	65,456,339	51%
	ITV10842	57,718,086	45%	ITV10844	78,135,173	48%
Mata	ITV10861	74,674,804	44%	ITV10863	70,326,864	47%
	ITV10862	46,994,527	43%	ITV10864	88,002,369	45%
Canga	ITV10891	75,377,499	43%	ITV10893	91,026,716	47%
	ITV10892	71,832,498	42%	ITV10894	55,450,224	49%
Estéril vermelho	ITV10911	61,143,181	44%	ITV10913	63,790,908	49%
	ITV10912	58,685,765	44%	ITV10914	67,296,521	53%

O programa utilizado para análise de controle de qualidade das sequências gerou um relatório com várias informações úteis para a avaliação do conjunto de dados gerado no sequenciamento das amostras, dentre eles: o valor de qualidade por base das *reads*; valor de qualidade médio das *reads*; o conteúdo de bases por posição nas *reads*; o conteúdo de GC por *read*; os níveis de duplicação das *reads*; as sequências super-representadas; o conteúdo de adaptadores; e o conteúdo de k-mers.

A média de qualidade por base das sequências foi em torno de PHRED 30. No final das *reads* ocorreu uma queda nesse valor, ficando abaixo de PHRED 20. Em geral, um score de qualidade é considerado aceitável quando é superior ao valor 20 na escala PHRED, dessa forma, foi necessário trinar o final das sequências com base nesse valor de qualidade. O valor de qualidade médio das sequências foi em torno de 32, mas algumas *reads* tiveram um valor de qualidade inferior a 20. Assim, também foi preciso descartar as sequências de baixa qualidade.

O modelo de distribuição do conteúdo GC por sequência é uma curva normal onde o valor central correspondente ao teor global médio de GC do genoma. Observamos que a distribuição real do conteúdo GC por *read* possuiu alguns picos de maior conteúdo GC, indicando a presença de sequências super-representadas ou de algum contaminante específico, como adaptadores. Foi observado sequências super-representadas nas bibliotecas das duas espécies bem como a presença de adaptadores universais da Illumina. Ao realizar BLAST das sequências super-representadas obtivemos como resultado alta similaridade com sequências referentes a rRNAs e a genes do cloroplasto de outras espécies de *Parkia*. A presença de sequências super-representadas de cloroplasto encontrado nas bibliotecas era esperada devido ao alto número de cópias de genoma plastidial em plantas.

Após avaliação da qualidade das *reads* geradas no sequenciamento foi realizada a retirada de sequências de adaptadores, de *reads* correspondentes a rRNAs e o tratamento para remoção das bases de baixa qualidade e das sequências menores que 35 pb. Em média 95% e 86% das *reads* provenientes das bibliotecas de *P. platycephala* e *S. pulcherrimum*, respectivamente, permaneceram após o corte (Tabela 6).

**Tabela 6:** Resultado do processamento das *reads* por qualidade e remoção de rRNA. PE = Os dois pares das *reads* paired-end permaneceram após o corte. SE = Apenas um par (single-end) das *reads* permaneceu após o corte.

<i>P. platycephala</i>				<i>S. pulcherrimum</i>			
Amostra	Número de reads inicial	Número de reads final (PE)	Número de reads final (SE)	Amostra	Número de reads inicial	Número de reads final (PE)	Número de reads final (SE)
ITV10841	67,271,933	33,427,112	30,735,642	ITV10843	65,456,339	22,825,371	32,802,578
ITV10842	57,718,086	29,950,013	25,153,855	ITV10844	78,135,173	36,668,004	29,675,539
ITV10861	74,674,804	40,889,774	30,382,345	ITV10863	70,326,864	29,754,556	32,218,505
ITV10862	46,994,527	23,914,183	20,983,469	ITV10864	88,002,369	39,519,928	41,400,803
ITV10891	75,377,499	38,493,332	33,339,698	ITV10893	91,026,716	40,025,333	40,059,362
ITV10892	71,832,498	36,016,468	32,651,956	ITV10894	55,450,224	23,407,709	23,885,672
ITV10911	61,143,181	29,859,628	28,211,003	ITV10913	63,790,908	26,961,220	27,626,019
ITV10912	58,685,765	31,611,250	24,343,221	ITV10914	67,296,521	24,366,861	30,253,546

### 4.1.3 – Montagem dos transcriptomas de referência.

As *reads* de alta qualidade advindas das bibliotecas sequenciadas foram utilizadas para a montagem dos transcritos das duas espécies. Nenhuma ferramenta de montagem de transcriptoma *de novo* fornece os melhores resultados para todos os conjuntos de dados (LU et al., 2013; GILBERT, 2013; VENTURINI et al., 2018). Vários trabalhos comparam a performance de diferentes softwares, gerando várias montagens diferentes, e selecionam uma como sendo a melhor para aquele organismo (GARG et al., 2011; ZHANG et al., 2013; CHAG et al., 2014). No entanto, cada montador e conjunto de parâmetros é melhor em algum ponto (recuperar isoformas, múltiplos tamanhos de k-mer, transcritos longos), produzindo alguns transcritos melhores que os outros (NAKASUGI et al., 2014). Nesse trabalho, quatro montadores *de novo*, implementando vários tamanhos de k-mer, foram utilizados para gerar várias montagens e selecionar o melhor conjunto de transcritos possível, conforme recomendado por Gilbert (2013). No total, 15 montagens foram geradas para *P. platycephala* e 17 para *S. pulcherrimum* (Tabela 7). Todas as montagens foram posteriormente concatenadas em uma única montagem combinada para cada espécie. Para ambas as espécies, Trinity produziu montagens com o maior comprimento médio de transcritos, maior valor de N50 e maior número de bases montadas. SOAPDenovo-Trans produziu os transcritos mais longos (37.652 pb com tamanho de k-mer 61 para *P. platycephala* e 49.517 pb com tamanho de k-mer 27 para *S. pulcherrimum*), mas também o maior número de transcritos para ambas as espécies. O número de contigs variou de 121.698 e 94.733 com Velvet / Oases usando um tamanho de k-mer de 61 a 255.644 e 233.410 com SOAPDenovo-Trans e um k-mer de 31 para *P. platycephala* e *S. pulcherrimum*, respectivamente.

A abordagem de montagem combinada permitiu a recuperação de muitos transcritos potenciais e suas variantes (Tabela 7), mas resultou em montagens muito grandes contendo muitas sequências redundantes (Tabela 8). As montagens combinadas de *P. platycephala* e *S. pulcherrimum* incluíram 2.794.335 e 2.898.193 transcritos montados *de novo*, respectivamente (Tabela 7). As montagens foram reduzidas com o pipeline do EvidentialGene e resultaram em 405.001 e 393.111 transcritos para *P. platycephala* e *S. pulcherrimum*, com 84.422 e 85.521 classificados como o conjunto principal de transcritos (transcritos primários), respectivamente.

O número de transcritos resultante do EvidentialGene para *P. platycephala* e *S. pulcherrimum* foi 7x maior que o número de transcritos encontrado para outras espécies de Caesalpinioideae (TORALES et al., 2013). A maioria dos transcritos é referente a proteínas curtas não classificadas. 28.561 (*P. platycephala*) e 27.632 (*S. pulcherrimum*) dos transcritos

primários codificam proteínas de 120 aminoácidos (120aa) ou mais. De acordo com Gilbert (2019), a maioria das proteínas putativas curtas (menores de 120aa) são loci espúrios que podem ser descartados se nenhuma evidência adicional de classificação for estabelecida. Dessa forma, foi feito um BLAST (e-value  $1 \times 10^{-6}$ ) das 55.861 (*P. platycephala*) e 57.889 (*S. pulcherrimum*) proteínas curtas preditas dos transcritos primários contra o banco de dados de Viridiplantae (UniProtKB / Swiss-Prot) na tentativa de estabelecer evidências de homologia para elas. Apenas 3.167 e 3.679 proteínas curtas, para *P. platycephala* e *S. pulcherrimum* respectivamente, tiveram correspondência com o banco de dados de Viridiplantae e foram mantidas para as análises posteriores. Os transcritos referentes a essas proteínas juntamente com os outros transcritos primários (cujas proteínas preditas possuem mais 120aa) serão referidos a partir daqui como os transcritos primários filtrados. As proteínas curtas restantes e suas formas alternativas foram descartadas. A montagem final de cada espécie consiste dos transcritos primários filtrados adicionados às suas formas alternativas e foram usadas para as análises de expressão diferencial espécie-específica. Os transcritos primários filtrados foram usados para a previsão de ortologia e para as análises de expressão diferencial dos genes ortólogos.

Tabela 7: Resumo das montagens.

Software	K-mer	<i>Parkia platycephala</i>						<i>Stryphnodendron pulcherrimum</i>					
		Nº de transcritos	Nº de bases	N50	Tamanho médio	Transcrito mais longo	Conteúdo GC (%)	Nº de transcritos	Nº de bases	N50	Tamanho médio	Transcrito mais longo	Conteúdo GC (%)
Trinity	25	170.619	265.016.620	2.690	1.553,27	22.754	39,81	163.406	170.691.286	1.883	1.044,58	21.057	40,45
rnaSpades	73	138.139	160.365.101	2.052	1.160,897	22.410	39,34	135.921	117.527.463	1.583	864,675	30.003	39,8
SoapDenovo-Trans	21							196.455	103.834.637	1.273	528	29.010	39,65
	27							217.812	110.135.287	1.278	505	49.517	39,75
	31	255.644	159.292.476	2.070	623,103	29.463	39,38	233.410	112.020.745	1.215	479	39.612	39,83
	39							229.080	111.464.951	1.164	486	28.310	39,95
	41	239.160	157.082.961	1.953	656	29.572	39,39	223.530	110.626.289	1.163	494	49.090	39,97
	51	248.106	154.263.042	1.803	621	29.372	39,42	222.797	108.155.163	1.128	485	37.559	40,2
	57	240.508	150.016.663	1.725	623	21.081	39,44						
	61	227.520	145.835.347	1.680	640	37.652	39,47	192.499	101.225.356	1.160	525	39.181	40,41
	71	195.307	132.050.477	1.560	676	28.775	39,56	155.743	90.382.458	1.213	580	39.640	40,7
Velvet/Oases	81	166.176	116.790.614	1.449	702	26.317	39,71						
	21	181.126	67.771.769	488	374,17	8.711	39,38	153.578	53.084.883	432	345,65	10.226	40,06
	27							149.929	50.907.620	419	339,54	9.129	39,99
	31	177.983	63.247.202	446	355,36	7.993	39,3	148.263	49.710.392	412	335,29	9.082	40,05
	39							136.060	47.610.109	423	349,92	10.178	40,26
	41	161.665	59.563.578	448	368,44	9.525	39,31	131.979	46.934.166	426	355,62	6.703	40,26
	51	141.349	56.327.081	468	398,5	6.989	39,28	112.998	43.729.053	450	386,99	7.366	40,52
Montagem combinada	57	129.335	54.306.642	485	419,89	8.777	39,29						
	61	121.698	53.155.197	501	436,78	7.669	39,36	94.733	40.873.130	496	431,46	5.797	40,88
Montagem pós EvidentalGene		2.794.335	1.795.084.770	1.455	642,401	37.652	38,85	2.898.193	1468.912.988	915	506,838	49.517	39,57
Montagem final		405.001	524.402.878	2.234	1.294,819	22.019	40,06	393.111	414.563.179	1.824	1.054,57	39.181	40,92
Primários final		271.668	453.536.490	2.488	1.669,451	22.019	40,47	247.694	347.927.703	2.098	1.404,667	39.181	41,35
		31.728	54.125.185	2.324	1.705,91	21.224	41,19	31.311	47.369.527	2.050	1.512,87	39.181	41,94

#### 4.1.4 – Verificação de qualidade e predição de ortólogos.

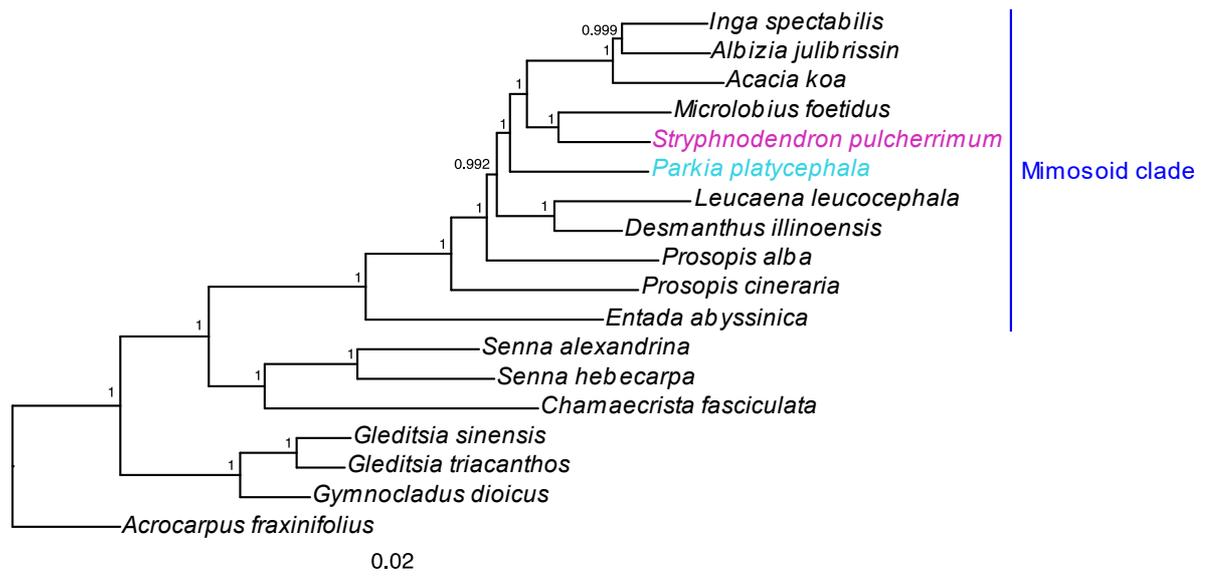
A qualidade e integridade do transcriptoma podem ter um impacto substancial na anotação e nas análises posteriores. Erros na montagem podem afetar a previsão de ortologia, o sinal filogenético e a quantificação da expressão gênica (HSIEH et al., 2019; SPILLANE et al., 2020). Uma métrica para avaliação é a representação das *reads* utilizadas como *input* na montagem. Uma montagem de alta qualidade deve representar a maior parte das *reads* utilizadas como *input*. Ao alinhar as *reads* de RNA-Seq de volta às montagens dos respectivos transcriptomas foi obtida uma taxa de alinhamento total de 96,78% para *P. platycephala* e 98,18% para *S. pulcherrimum*, sendo que 95,14% e 96,52% dos pares alinharam concordantemente no mesmo transcrito, o que é indicativo de uma montagem não fragmentada.

Outra avaliação importante da montagem é a completude e integridade dos genes ortólogos presentes. O resultado do BUSCO com os dois bancos de dados utilizados está representado na Tabela 8. Mais de 94% do total de genes ortólogos de cópia única de embriófitas e eudicotiledôneas buscados nas montagens foram encontrados completos. Grande parte desses genes foram encontrados duplicados nas montagens, mesmo após a redução do EvidentialGene, devido à presença de múltiplas isoformas. A análise do BUSCO com os transcritos primário filtrados (que exclui isoformas alternativas) das duas espécies mostrou uma diminuição acentuada dos genes duplicados (de aproximadamente 80% para aproximadamente 8% e 5%, para *P. platycephala* e *S. pulcherrimum*, respectivamente), mantendo a recuperação de ortólogos completos de embriófitas e eudicotiledôneas acima de 90% (Tabela 8). Além disso, apenas entre 1,9% e 7,2% dos genes quase universais foram classificados como ausentes nos transcriptomas das plantas em todas as análises, indicando alta qualidade e boa cobertura dos transcriptomas.

**Tabela 8:** Resultados do BUSCO com os transcriptomas de novo de *P. platycephala* e *S. pulcherrimum*. Os transcriptomas de *P. platycephala* e *S. pulcherrimum* foram comparados com os bancos de dados de embriófitas (linhas superiores) e eudicotiledôneas (linhas inferiores). A porcentagem de ortólogos inferidos está entre parênteses ao lado do número de ortólogos.

BUSCOs	<i>Parkia platycephala</i>			<i>Stryphnodendron pulcherrimum</i>		
	Montagem combinada	Montagem pós EvidentialGene	Primários filtrados	Montagem combinada	Montagem pós EvidentialGene	Primários filtrados
Completo	1358 (94.3%)	1357 (94.3%)	1314 (91.2%)	1355 (94.1%)	1354 (94.0%)	1312 (91.1%)
	2041 (96.2%)	2046 (96.5%)	2003 (94.5%)	2027 (95.6%)	2037 (96.0%)	1981 (93.4%)
Completo e cópia única	50 (3.5%)	214 (14.9%)	1201 (83.4%)	34 (2.4%)	205 (14.2%)	1236 (85.8%)
	70 (3.3%)	311 (14.7%)	1840 (86.8%)	47 (2.2%)	312 (14.7%)	1862 (87.8%)
Completo e duplicado	1308 (90.8%)	1143 (79.4%)	113 (7.8%)	1321 (91.7%)	1149 (79.8%)	76 (5.3%)
	1971 (92.9%)	1735 (81.8%)	163 (7.7%)	1980 (93.4%)	1725 (81.3%)	119 (5.6%)
Fragmentado	28 (1.9%)	28 (1.9%)	44 (3.1%)	22 (1.5%)	21 (1.5%)	24 (1.7%)
	41 (1.9%)	34 (1.6%)	52 (2.5%)	39 (1.8%)	31 (1.5%)	52 (2.5%)
Faltando	54 (3.8%)	55 (3.8%)	82 (5.7%)	63 (4.4%)	65 (4.5%)	104 (7.2%)
	39 (1.9%)	41 (1.9%)	66 (3.0%)	55 (2.6%)	53 (2.5%)	88 (4.1%)

Para confirmar as relações entre as duas espécies estudadas bem como a acurácia e resolução dos transcriptomas de referência, uma filogenia molecular da subfamília Caesalpinioideae foi construída a partir dos ortogrupos de cópia única entre as duas espécies e outras 16 espécies da subfamília, inferidos pelo OrthoFinder. Um ortogrupo é definido como um grupo de genes descendentes de um único gene do último ancestral comum dentro de um grupo de espécies (EMMS e KELLY, 2015). Como esperado, o filograma dos ortólogos mostra o posicionamento das duas espécies dentro do clado Mimosoidea (Figura 7). Nesse estudo, ambas as espécies fazem parte de um subclado composto pelos gêneros *Inga*, *Albizia*, *Acacia* e *Microlobius*. *P. platycephala* foi a primeira espécie a divergir deste subclado e é irmã das demais espécies. *S. pulcherrimum* foi colocada como irmã de *Microlobius foetidus*, conforme inferido a partir de dados de cloroplasto e ITS (SIMON et al., 2016). Filogenias recentes da família Fabaceae ou de clados dentro da família, com base em dados nucleares e plastidiais (SIMON et al., 2016; WANG et al., 2017; LPWG, 2017; KOENEN et al., 2020) corroboram a topologia de Caesalpinioideae encontrada neste estudo, indicando a consistência filogenética da subfamília e a qualidade dos transcriptomas montados.



**Figura 7:** Árvore de máxima verossimilhança aproximada baseada em ortólogos de cópia única de espécies de Caesalpinioideae, gerada com OrthoFinder/FastTree. Os números nos pontos de ramificação representam valores de suporte baseados em 1.000 reamostragens de bootstrap local (teste de Shimodaira-Hasegawa) geradas com FastTree.

As informações inferidas pelo OrthoFinder também foram utilizadas para identificar os genes ortólogos de cópia única entre as duas espécies estudadas. Das 63.039 proteínas preditas de ambas as espécies combinadas, um total de 53.763 foi atribuído a 16.958 ortogrupos. 16.010 (94,4%) dos ortogrupos foram compartilhados, 948 foram classificados como espécie-específicos e 9.284 eram de cópia única para ambas as espécies. Os ortogrupos de cópia única compartilhados foram usados para as análises DE dos ortólogos.

## 4.2 Plasticidade da expressão gênica nos substratos naturais

### 4.2.1 – Sobre o número de réplicas biológicas.

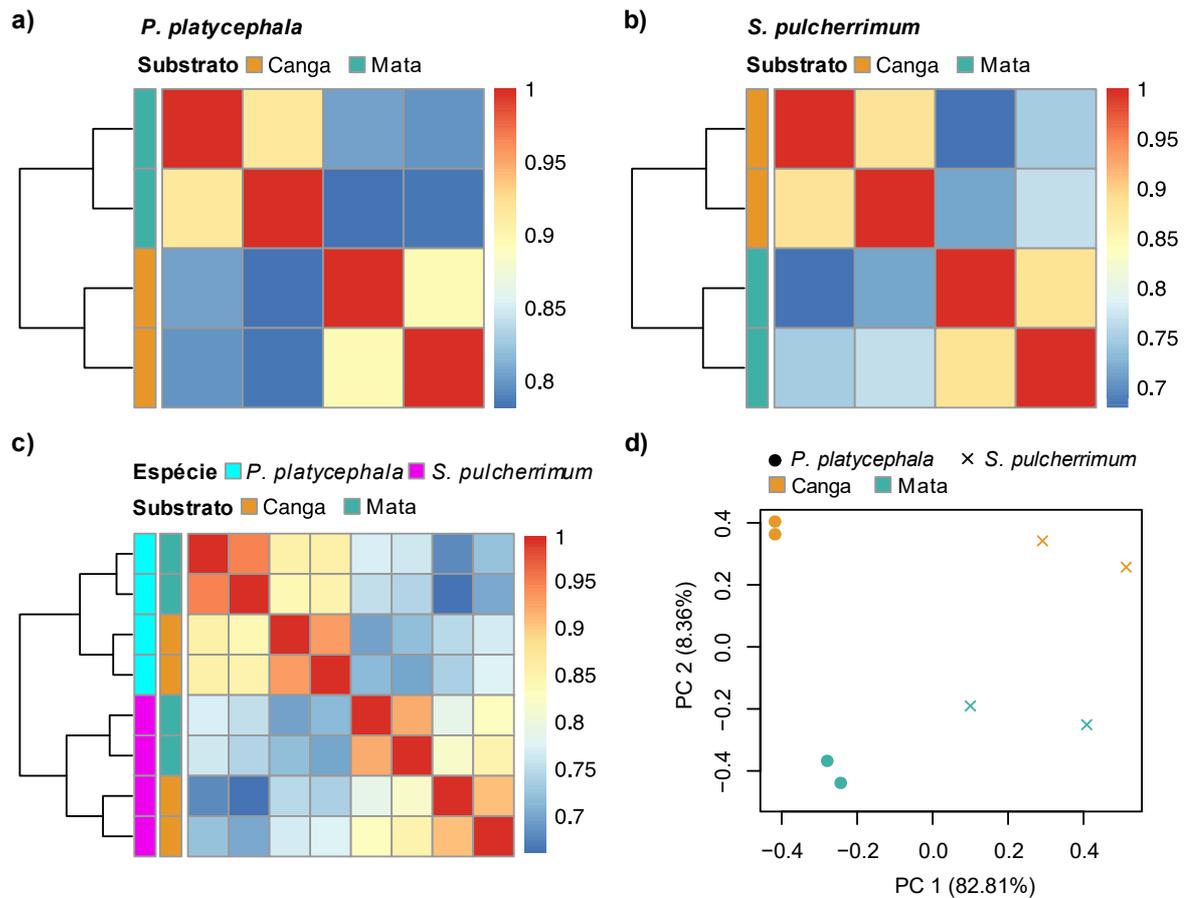
As amostras foliares de três indivíduos cultivados na mesma bandeja foram agrupadas para extração de RNA e sequenciamento. Amostras provenientes de duas bandejas por substrato foram sequenciadas, gerando duas réplicas biológicas por condição. Com apenas duas réplicas, o poder de nossas estimativas na identificação de GDEs é reduzido. Ainda assim, o edgeR identifica com sucesso a maior parte dos GDEs verdadeiros (>85%) com um número extremamente baixo de falsos positivos, para os genes mais fortemente alterados ( $|\log_2FC| \geq 2$ ), mesmo com apenas duas réplicas por condição (RAPAPORT et al., 2013; SCHURCH et al., 2016) e novas hipóteses podem ser levantadas. Dessa forma, utilizamos um corte rigoroso de

$|\log_2FC| = 2$  e  $FDR = 0,001$  para garantir uma alta taxa de verdadeiros positivos encontrados. Embora isso possa aumentar as taxas de falsos negativos, fortalece nossa confiança de que resultados significativos revelam os efeitos do tratamento. Além disso, o agrupamento reduz a variabilidade biológica entre as amostras, aumentando as chances de nossos resultados corresponderem à variabilidade entre os grupos e permite a detecção de efeitos biológicos com um número pequeno de amostras (ASSEFA et al., 2020).

#### 4.2.2 – Relação entre as amostras

Para avaliar a relação entre as amostras, os perfis de expressão gênica foram investigados com uma análise de agrupamento hierárquico com base nos coeficientes de correlação de *Spearman* dos valores de expressão gênica normalizados entre pares de amostras. As análises foram realizadas para cada espécie separadamente, considerando todos os genes minimamente expressos em cada uma e para as duas espécies combinadas, considerando os valores de expressão dos ortólogos de cópia única compartilhados entre elas.

Os mapas de calor indicam boa consistência entre as amostras de cada condição, mostrando correlações mais altas para a mesma condição do que entre as condições para ambas as espécies (Figura 8a-b). O mapa de calor combinando os valores de expressão dos genes ortólogos nas duas espécies identificou padrões semelhantes, no entanto, as diferenças entre as espécies são maiores do que entre as condições de crescimento das plantas, sugerindo um padrão geral de expressão espécie-específico (Figura 8c). A análise de componentes principais (*principal component analysis* - PCA) também evidenciou um padrão de expressão espécie-específico, com o primeiro componente (PC1) separando claramente as duas espécies (Figura 8d). As duas condições (canga e mata) foram separadas no segundo componente (PC2) (Figura 8d).



**Figura 8:** Perfil de expressão gênica de *P. platycephala* e *S. pulcherrimum* no desenvolvimento nos substratos naturais. a) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras de *P. platycephala* considerando todos os genes expressos. Os coeficientes de correlação são mostrados à direita variando de azul (baixa correlação) a vermelho (alta correlação) b) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras de *S. pulcherrimum* considerando todos os genes expressos. c) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras das duas espécies considerando os genes ortólogos cópia única compartilhados. d) PCA dos níveis de expressão normalizados dos ortólogos de cópia única nas duas espécies. As espécies são representadas pelo formato dos pontos. Os substratos são representados pela cor dos pontos.

#### 4.2.3 – Análises de expressão diferencial

Para caracterizar os padrões de plasticidade no transcriptoma das espécies como um todo, GDEs foram identificados em cada espécie separadamente, comparando os dados de expressão gênica em plantas cultivadas em substrato de mata (controle) com aquelas cultivadas em substrato de canga.

Os GDEs foram identificados usando o pacote edgeR. Os resultados das análises espécie-específicas indicam que apenas uma minoria dos genes em ambas as espécies respondeu plasticamente à diferença de substrato (aproximadamente 5% do total de genes

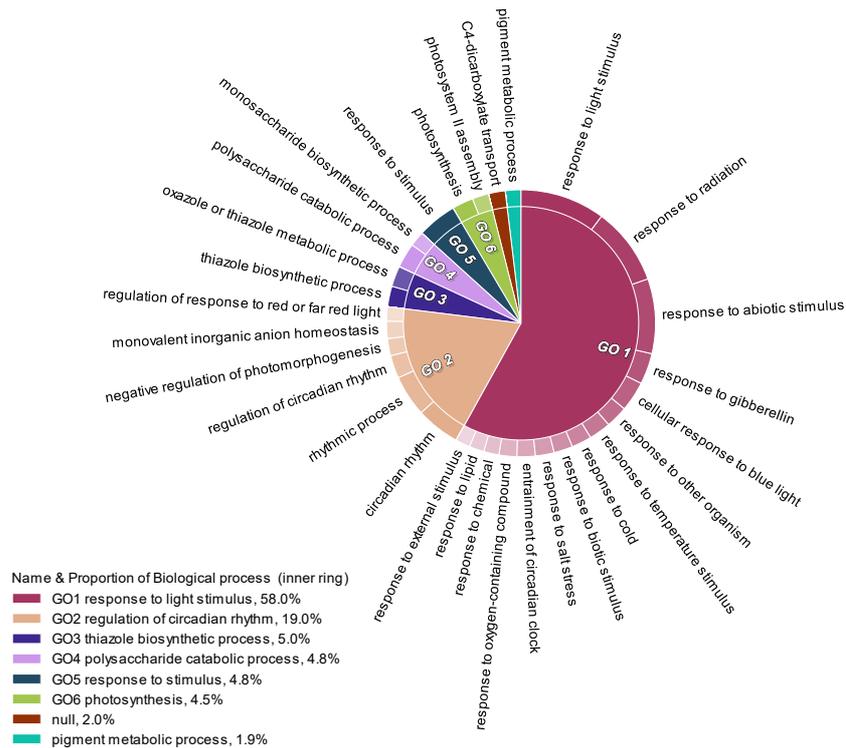
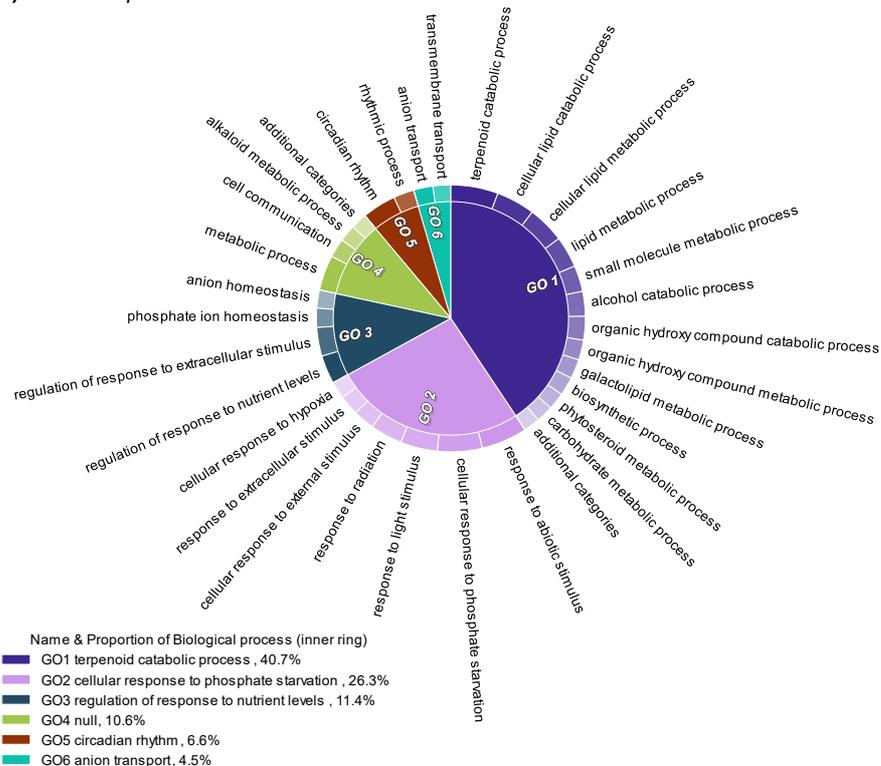
analisados para cada uma - Tabela 9). Ambas as espécies tiveram mais genes regulados negativamente ao comparar plantas cultivadas no substrato de canga com plantas cultivadas em substrato de mata.

**Tabela 9:** Plasticidade da expressão gênica das plantas no desenvolvimento em substratos de canga e mata para todos os genes e para os pares de genes ortólogos de *P. platycephala* e *S. pulcherrimum*. Total – número de genes inicial; > 1 cpm – número de genes utilizados para as análises após filtragem por nível de expressão, cpm = count per million; DE canga x mata – Número de GDEs nas plantas cultivadas em canga versus plantas cultivadas em mata; Up – Número de genes regulados positivamente; Down – Número de genes regulados negativamente.

	Número de genes		
	<i>P. platycephala</i>	<i>S. pulcherrimum</i>	Ortogrupos
Total	31.731	31.317	9284
> 1 cpm	19.927	16.669	8220
DE canga x mata	1.112	838	298
Up	390	264	94
Down	723	574	203

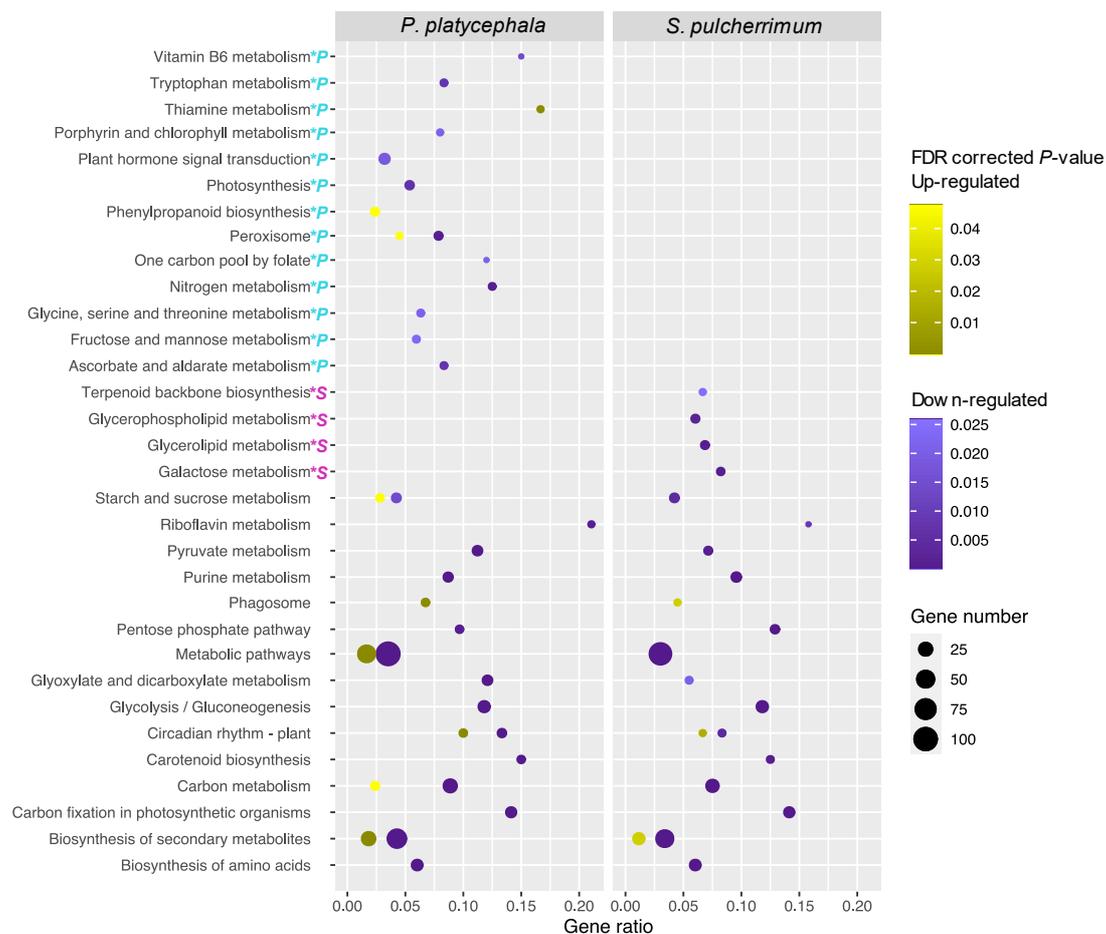
Para classificar a função biológica dos GDEs, foi feita a investigação dos termos GO enriquecidos usando Goseq (YOUNG et al., 2010). A análise de enriquecimento dos genes super-expressos identificou 28 termos enriquecidos ( $FDR < 0,05$ ) para *P. platycephala* para as categorias processo biológico (BP), componente celular (CC) e função molecular (MF). Para *S. pulcherrimum*, apenas o termo “processo rítmico” foi encontrado enriquecido nos genes super-expressos. Para os genes sub-expressos, foram identificados 85 (*P. platycephala*) e 131 (*S. pulcherrimum*) termos GO enriquecidos para as categorias BP, CC e MF (Anexos 1 e 2).

Para *P. platycephala*, dentro da categoria processo biológico, os GDEs enriquecidos foram associados principalmente à resposta a estímulo abiótico. Os termos GO "processo rítmico", "ritmo circadiano", resposta a estímulos (abióticos, externos, luminoso, temperatura) foram significativamente enriquecidos tanto nos genes super-expressos quanto nos sub-expressos. No entanto, os termos GO "processo biossintético de oxazol ou tiazol", "processo catabólico de polissacarídeo" e "resposta ao estímulo biótico" foram enriquecidos apenas nos genes super-expressos. Os termos “resposta à giberelina” e “fotossíntese” foram enriquecidos apenas nos genes sub-expressos (Anexo 1). Para *S. pulcherrimum*, o termo GO "processo rítmico" foi enriquecido tanto nos genes super-expressos quanto nos sub-expressos. Os genes sub-expressos estão associados a vários processos catabólicos (terpenos, lipídios, ácido abscísico), à resposta a vários estímulos (abióticos, luminoso, radiação, nível de nutrientes), e "resposta celular à deficiência de fosfato" (Anexo 2). Os termos GO, pertencentes a categoria BP, foram reduzidos para diminuição da redundância com o REVIGO e podem ser visualizados na Figura 9.

a) *P. platycephala*b) *S. pulcherrimum*

**Figura 9:** 15 principais processos biológicos enriquecidos ( $p < 0,01$ ) nos GDEs para (a) *P. platycephala* e (b) *S. pulcherrimum*. Os gráficos foram construídos com CirGO em uma estrutura hierárquica de dois níveis. A legenda mostra os rótulos “pais”, representados no anel interno do gráfico de pizza, e a proporção da fatia. O anel externo do gráfico de pizza representa a contribuição relativa dos rótulos “filhos”. Os nomes dos termos foram mantidos no idioma original.

Para caracterizar ainda mais a função dos GDEs, a análise das vias metabólicas enriquecidas foi realizada usando o banco de dados KEGG (*Kyoto Encyclopedia of Genes and Genomes* - <https://www.genome.jp/kegg/>) com o KOBAS (MAO et al., 2005). Identificamos 28 vias enriquecidas (FDR <0,05) em *P. platycephala* e 19 em *S. pulcherrimum*, 15 das quais em comum entre as espécies. A Figura 10 mostra todas as vias KEGG enriquecidas nos GDEs para ambas as espécies. As vias enriquecidas comuns às duas espécies estão relacionadas a biossíntese de metabólitos secundários e dos seus precursores no metabolismo dos carboidratos e na biossíntese de aminoácidos. Embora ambas as espécies apresentem vias alteradas de forma semelhante, vias metabólicas alteradas exclusivamente em uma das espécies também foram encontradas. Para *P. platycephala*, essas vias estão relacionadas ao metabolismo cofatores e vitaminas do complexo B (porfirina e clorofila; tiamina; vitamina B6), biossíntese de fenilpropanoides, fotossíntese e peroxissomo. Para *S. pulcherrimum*, as vias exclusivas estão relacionadas ao metabolismo de lipídios (glicerolipídios; glicerofosfolipídios) e metabolismo de terpenóides.



**Figura 10:** Gráfico de dispersão das vias KEGG enriquecidas em cada espécie cultivadas em substratos de canga em comparação com substratos de mata. Esquerda - *P. platycephala*; Direita - *S. pulcherrimum*. As vias KEGG significativamente enriquecidas (FDR <0,05, teste exato de Fisher) foram indicadas

como pontos. Os tamanhos dos pontos representam o número de genes incluídos em cada via. O eixo x representa a razão entre o número de genes expressos diferencialmente e o número de todos os genes na via. Genes super-expressos foram coloridos em tons de amarelo; genes sub-expressos em tons de roxo. Quanto mais profunda for a cor, menor o valor de FDR. As letras coloridas ao lado dos nomes das vias marcam vias exclusivamente alteradas em uma das espécies. *P* em ciano - *P. platycephala*; *S* em rosa - *S. pulcherrimum*. Os nomes das vias foram mantidos no idioma original.

Para facilitar a comparação das respostas entre as duas espécies e observar a existência de genes que são alterados da mesma maneira no desenvolvimento inicial das espécies na canga, as informações de ortologia gênica de ambas as espécies inferidas pelo OrthoFinder foram utilizadas. Após corrigir a diferença de expressão gênica basal entre as duas espécies, foi feita a identificação dos ortólogos diferencialmente expressos entre o crescimento no substrato de canga e de mata com resposta similar para as duas espécies (ODEsRS). O perfil dos ODEsRS encontrados foi similar ao das análises das espécies separadas, com um maior número de ortólogos sub-expressos (Tabela 9). A análise de enriquecimento mostrou 26 termos GO enriquecidos (FDR <0,05) (Tabela 10). A maioria dos termos na categoria processo biológico está relacionada à regulação da transcrição, ao ritmo circadiano e à resposta a luminosidade. A análise de enriquecimento das vias metabólicas mostrou apenas duas vias enriquecidas (FDR <0,05) nos ODEsRS: Ritmo circadiano e metabolismo de riboflavina. Esses genes podem estar relacionados ao sucesso de estabelecimento das espécies em ambientes variados.

**Tabela 10:** Termos Gene Ontology (GO) identificados como enriquecidos pelo teste exato de Fisher entre os ODEsRS nas plantas cultivadas na canga. O nome dos termos foi mantido no idioma original.

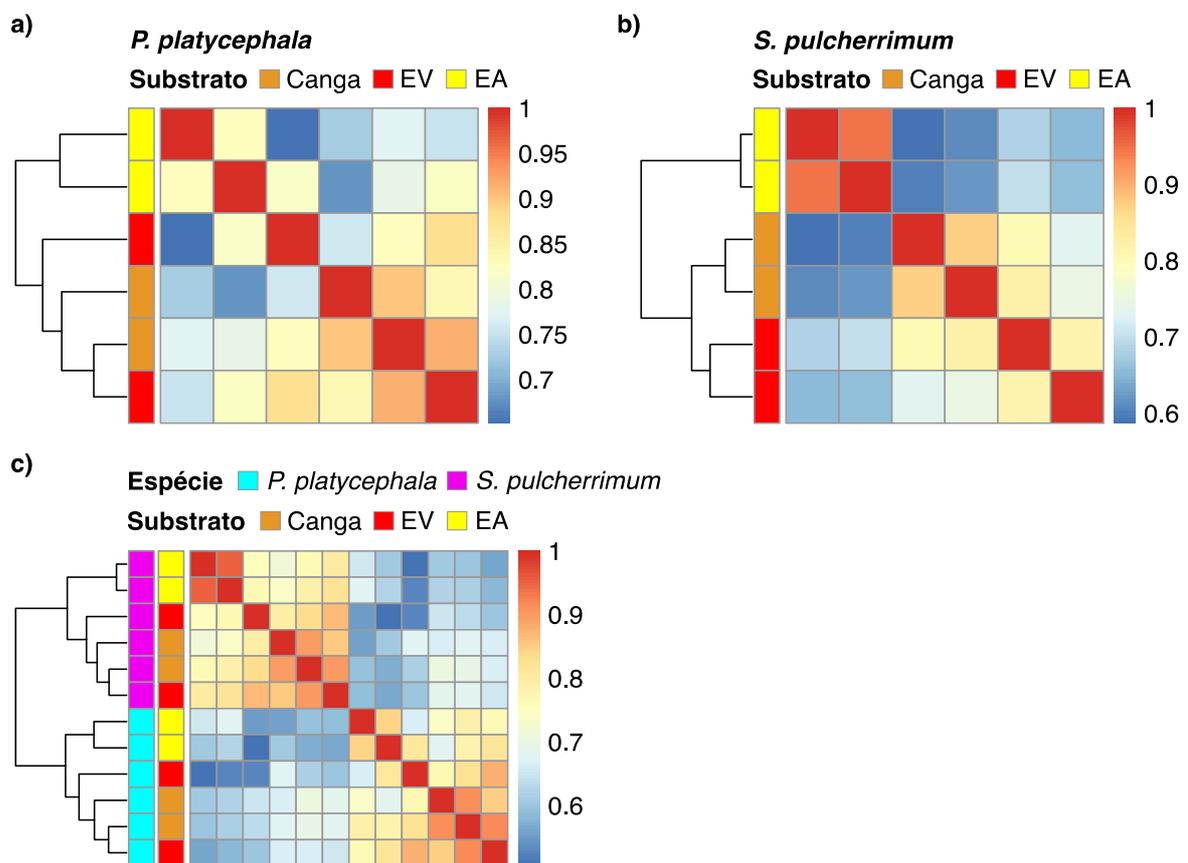
Acesso GO	Categoria	Termo	FDR	Nº de genes
GO:0009507	CC	chloroplast	1.51E-13	92
GO:0006355	BP	regulation of transcription, DNA-templated	2.76E-08	39
GO:0007623	BP	circadian rhythm	5.32E-08	12
GO:0010114	BP	response to red light	2.69E-07	9
GO:0005515	MF	protein binding	1.48E-06	60
GO:0003700	MF	DNA-binding transcription factor activity	6.36E-06	36
GO:0009649	BP	entrainment of circadian clock	1.64E-05	5
GO:0010017	BP	red or far-red light signaling pathway	1.64E-05	6
GO:0009535	CC	chloroplast thylakoid membrane	2.62E-05	14
GO:0009570	CC	chloroplast stroma	3.02E-05	19
GO:0009637	BP	response to blue light	0.00026005	6
GO:0080167	BP	response to karrikin	0.00036018	8
GO:0009416	BP	response to light stimulus	0.00053656	10
GO:0009658	BP	chloroplast organization	0.00135533	8
GO:0090351	BP	seedling development	0.00239464	4
GO:0000427	CC	plastid-encoded plastid RNA polymerase complex	0.00340051	3
GO:0032922	BP	circadian regulation of gene expression	0.00340051	3
GO:0042752	BP	regulation of circadian rhythm	0.00563682	4
GO:0042753	BP	positive regulation of circadian rhythm	0.006172	3
GO:0010218	BP	response to far red light	0.01466872	4
GO:0045892	BP	negative regulation of transcription, DNA-templated	0.01476035	7
GO:0009909	BP	regulation of flower development	0.01476035	5
GO:0009579	CC	thylakoid	0.02513588	7
GO:0009654	CC	photosystem II oxygen evolving complex	0.02601454	3
GO:0009534	CC	chloroplast thylakoid	0.03019725	7
GO:2000028	BP	regulation of photoperiodism, flowering	0.04020657	3

### 4.3 Respostas transcriptômicas ao crescimento em substratos de estéril de mina

#### 4.3.1 – Relação entre as amostras

A relação entre as amostras de plantas cultivadas nos substratos de canga e dos resíduos de mineração foram também avaliadas com uma análise de agrupamento hierárquico com base nos coeficientes de correlação de *Spearman* dos valores de expressão gênica normalizados. As análises foram realizadas para cada espécie separadamente, considerando todos os genes minimamente expressos em cada uma e para as duas espécies combinadas, considerando os valores de expressão dos ortólogos de cópia única compartilhados entre elas.

Nas análises espécies-específicas, as amostras de estéril amarelo se agruparam separadamente das amostras de canga e de estéril vermelho para as duas espécies (Figura 11a-b). Para *P. platycephala*, as amostras de canga e estéril vermelho se agruparam (Figura 11a). Esses resultados sugerem que o crescimento no substrato de estéril vermelho induz um perfil de expressão gênica semelhante ao crescimento no substrato da canga. Para *S. pulcherrimum*, as amostras se agruparam de acordo com o substrato de cultivo (Figura 11b). O mapa de calor combinando os valores de expressão dos genes ortólogos nas duas espécies novamente identificou padrões de expressão espécie-específicos (Figura 11c), como observado nas análises com os substratos naturais (Figura 8c). As amostras do substrato de estéril amarelo se agrupam separadamente das amostras de canga e estéril vermelho, como observado nas análises espécie-específicas. As amostras de canga e estéril vermelho se agruparam, novamente sugerindo que esses substratos induzem perfis de expressão gênica semelhantes.



**Figura 11:** Perfil de expressão gênica de *P. platycephala*, *S. pulcherrimum* no desenvolvimento nos substratos de canga e resíduos de mineração. EV – Estéril vermelho; EA – Estéril amarelo. a) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras de *P. platycephala* considerando todos os genes expressos. Os coeficientes de correlação são mostrados à direita variando de azul (baixa correlação) a vermelho (alta correlação). b) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras de *S. pulcherrimum* considerando todos os genes expressos. c) Mapa de calor simétrico dos coeficientes de correlação de *Spearman* entre as amostras das duas espécies considerando os genes ortólogos cópia única compartilhados.

### 4.3.2 – Análises de expressão diferencial

Com o objetivo de investigar a resposta transcriptômica no desenvolvimento inicial das espécies em substratos de áreas degradadas, além de avaliar se conseguimos prever o sucesso das espécies na ocupação das áreas, GDEs foram identificados em cada espécie após crescimento em canga e em dois tipos de substrato de estéril provenientes de áreas de mineração.

Os resultados das análises espécie-específicas novamente mostraram uma minoria de genes alterados no desafio de substrato (Tabela 11). Ambas as espécies tiveram um menor número de GDEs na comparação com o cultivo em estéril vermelho do que com o cultivo em estéril amarelo. O número de GDEs em estéril vermelho representa menos de 1% (aproximadamente 0,6%) dos genes minimamente expressos em ambas as espécies, indicando que o substrato de estéril vermelho pode não oferecer um desafio significativo no desenvolvimento das espécies, mesmo sendo mais ácido e nutricionalmente mais pobre do que o substrato de canga (Tabela 1; SILVA et al., 2018). O perfil de expressão das duas espécies em estéril vermelho foi semelhante, com um maior número de genes super-expressos.

O substrato de estéril amarelo apresenta um desafio maior para ambas as espécies por ser um substrato mais compacto, com menos matéria orgânica, dificultando o aporte de nutrientes. Ainda assim, *P. platycephala* parece ter conseguido compensar o estresse abiótico causado e teve um desenvolvimento inicial em estéril amarelo semelhante ao desenvolvimento em canga (SILVA et al., 2018). É interessante observar que o número de GDEs no crescimento em estéril amarelo foi 50% menor do que na comparação entre canga e mata para *P. platycephala*. A comparação entre esses dois grupos de GDEs deve diferenciar genes naturalmente plásticos na comparação de substratos de desenvolvimento e genes que estão diretamente relacionados a compensação de estresse. *S. pulcherrimum* teve proporcionalmente mais GDEs no substrato de estéril amarelo do que *P. platycephala* e a mesma proporção de GDEs na comparação entre os substratos naturais, evidenciando o maior desafio desse substrato para essa espécie.

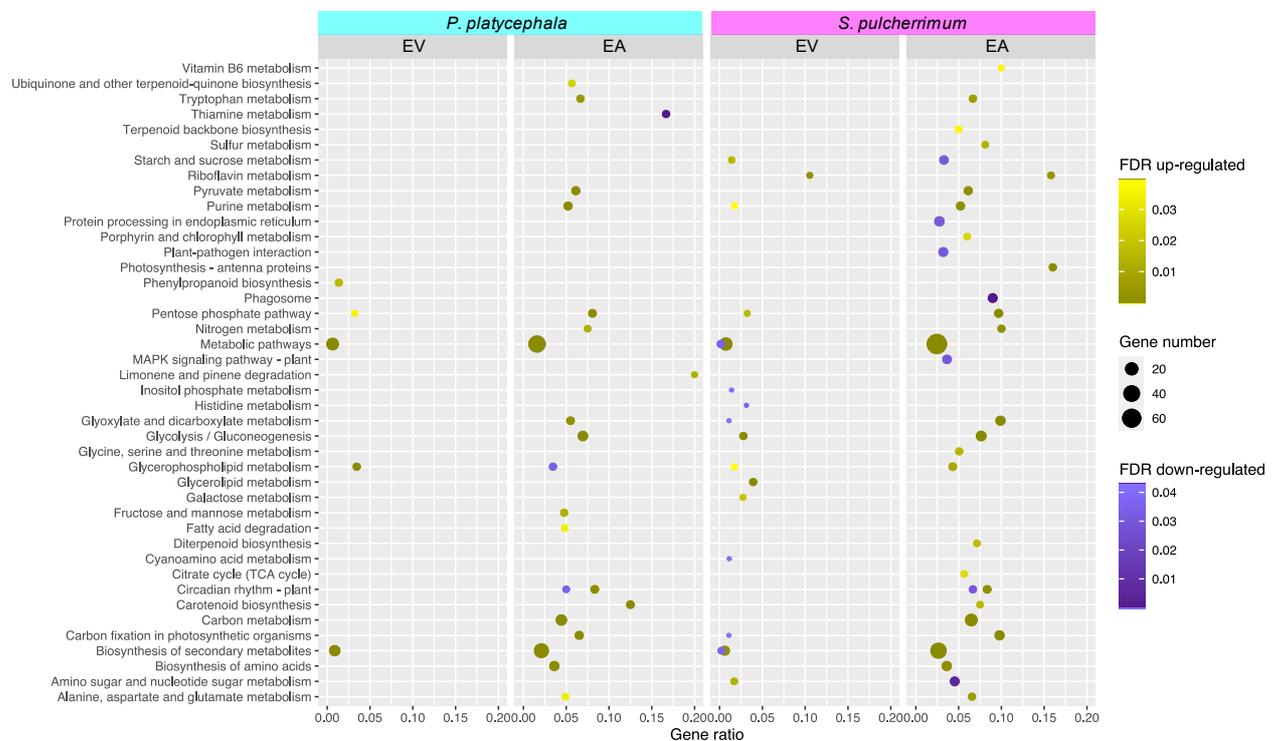
**Tabela 11:** Número de genes diferencialmente expressos no desenvolvimento em substratos de estéril de mina considerando todos os genes das espécies e para os pares de genes ortólogos de *P. platycephala* e *S. pulcherrimum*. Total – número de genes inicial; DEs EV x canga – Número de GDEs nas plantas cultivadas em estéril vermelho versus cultivadas em canga; DEs EA x canga – Número de GDEs nas plantas cultivadas em estéril amarelo versus cultivadas em canga; Up – Número de genes regulados positivamente; Down – Número de genes regulados negativamente.

	Número de genes		
	<i>P. platycephala</i>	<i>S. pulcherrimum</i>	Ortogrupos
DEs EV x canga	119	98	0
Up	102	79	
Down	17	19	
DEs EA x canga	603	881	267
Up	369	418	167
Down	234	463	100

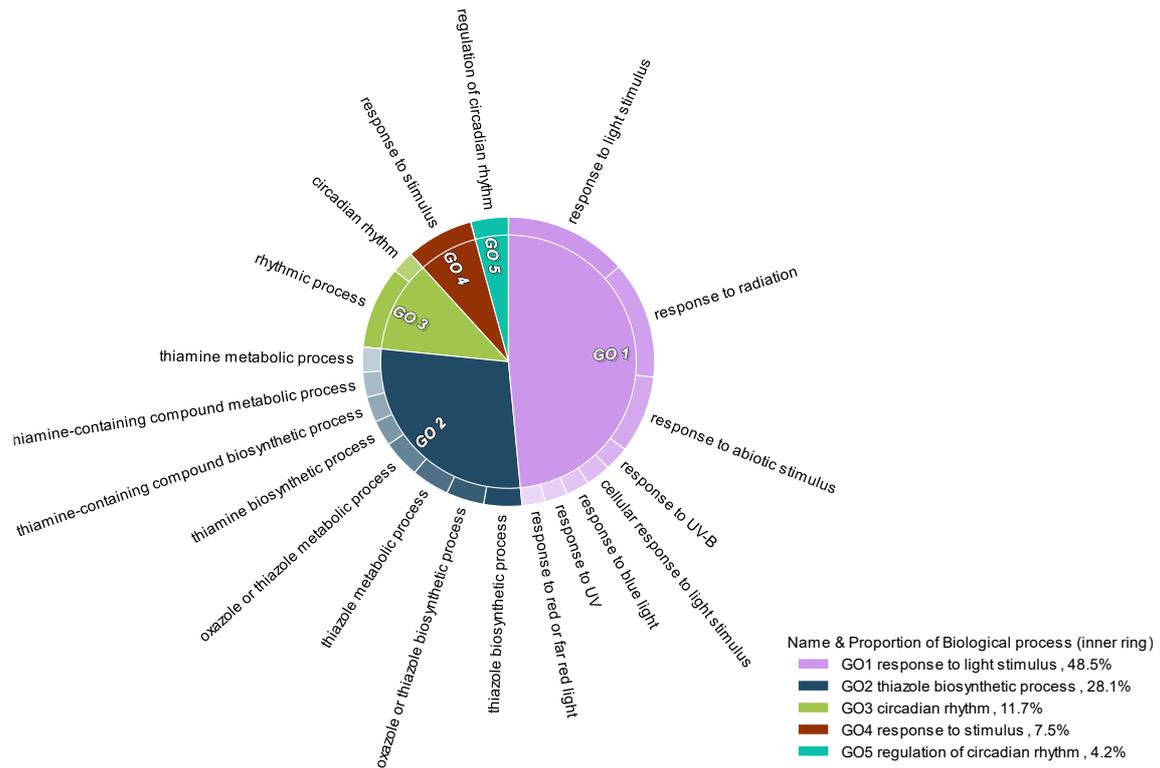
As análises de enriquecimento funcional dos GDEs nas plantas cultivadas em estéril vermelho não retornaram nenhum termo GO enriquecido para *P. platycephala*. Cinco vias metabólicas foram encontradas enriquecidas, todas nos genes super-expressos (Figura 12). Essas vias estão relacionadas à biossíntese de fenilpropanoides e metabolismo de glicerofosfolipídios. Nas plantas cultivadas em estéril amarelo, 30 termos GO foram encontrados enriquecidos em *P. platycephala* nas três categorias (Anexo 3), sendo 20 processos biológicos (Figura 13). Os termos relacionados ao ritmo circadiano foram enriquecidos tanto nos genes super-expressos quanto nos sub-expressos. Os termos relacionados à resposta a luz e radiação foram enriquecidos nos genes super-expressos e os termos relacionados ao metabolismo de tiamina e tiazol nos genes sub-expressos. 21 vias foram enriquecidas em *P. platycephala* em estéril amarelo, três nos genes sub-expressos (ritmo circadiano, metabolismo de tiamina e de glicerofosfolipídios) e 19 nos genes super-expressos, relacionadas à biossíntese de metabólitos secundários e dos seus precursores (Figura 12), semelhante à análise com os substratos naturais (Figura 10).

Os genes sub-expressos de *S. pulcherrimum* em estéril vermelho não exibiram termos GO enriquecidos. Os genes super-expressos apresentaram 46 termos enriquecidos, relacionados à resposta aos níveis de nutriente, à deficiência e homeostase de fosfato e resposta à queda nos níveis de oxigênio (Figura 14a; Anexo 4). 16 vias metabólicas foram enriquecidas, incluindo vias relacionadas ao metabolismo dos carboidratos, lipídios e aminoácidos, bem como de inositol fosfato (Figura 12). No substrato de estéril amarelo, os GDEs de *S. pulcherrimum* apresentaram 60 termos GO enriquecidos (19 nos genes sub-expressos e 41 nos super-expressos) nas três categorias (Anexo 5). Os termos da categoria processo biológico

enriquecidos nos genes sub-expressos estão relacionados a defesa, resposta imune e resposta a estímulo biótico. Nos genes super-expressos termos relacionados a resposta a estímulos abióticos (radiação, luz, nível de nutrientes), metabolismo de terpenóides, deficiência e homeostase de fosfato e resposta à queda nos níveis de oxigênio foram super-representados (Figura 14b). 31 vias metabólicas enriquecidas foram encontradas nos GDEs em estéril amarelo, relacionadas ao metabolismo energético (fotossíntese, metabolismo de nitrogênio e do enxofre, ciclo do ácido tricarboxílico [TCA]), ao metabolismo de terpenóides, interação planta-patógeno e vias de sinalização por MAPK (*Mitogen-Activated Protein Kinase*) (Figura 12).



**Figura 12:** Gráfico de dispersão das vias KEGG enriquecidas em cada espécie cultivadas em substratos de estéril de mina em comparação com substrato de canga. Esquerda - *P. platycephala*; Direita - *S. pulcherrimum*. EV – Estéril vermelho; EA – Estéril amarelo. As vias KEGG significativamente enriquecidas (FDR < 0,05, teste exato de Fisher) foram indicadas como pontos. Os tamanhos dos pontos representam o número de genes incluídos em cada via. O eixo x representa a razão entre o número de genes expressos diferencialmente e o número de todos os genes na via. Genes super-expressos foram coloridos em tons de amarelo; genes sub-expressos em tons de roxo. Quanto mais profunda for a cor, menor o valor de FDR.



**Figura 13:** Processos biológicos enriquecidos (FDR <0,05) nos GDEs para *P. platycephala* cultivada em estéril amarelo. O gráfico foi construído com CirGO em uma estrutura hierárquica de dois níveis. A legenda mostra os rótulos “pais”, representados no anel interno do gráfico de pizza, e a proporção da fatia. O anel externo do gráfico de pizza representa a contribuição relativa dos rótulos “filhos”. Os nomes dos termos foram mantidos no idioma original.



O número de GDEs no desenvolvimento em estéril vermelho foi pequeno para as duas espécies e a análise com os genes ortólogos compartilhados não retornou nenhum ODEsRS para essa condição. No desenvolvimento em estéril amarelo, 267 ODEsRS foram encontrados (Tabela 11). Nenhuma via metabólica enriquecida foi encontrada para esses genes. 51 termos GO foram encontrados enriquecidos, 29 nos genes sub-expressos relacionados a organização do cloroplasto, ritmo circadiano, resposta a luz e regulação da transcrição, e 26 nos super-expressos, também relacionados ao ritmo circadiano, resposta a luz e regulação da transcrição (Tabela 12). Esses resultados sugerem que as plantas alteram a expressão de genes relacionados a percepção da variação ambiental dependendo da composição do substrato e que esse efeito pode estar relacionado a alteração das funções relacionadas ao cloroplasto.

**Tabela 12:** Termos Gene Ontology (GO) identificados como enriquecidos pelo teste exato de Fisher entre os ODEsRS nas plantas cultivadas em estéril amarelo. O nome dos termos foi mantido no idioma original.

Acesso GO	Categoria	Termo	FDR	Nº de genes	GDEs
GO:0003700	MF	DNA-binding transcription factor activity	1.82E-09	24	Up
GO:0005515	MF	protein binding	1.75E-09	31 / 15	Up/ Down
GO:0007623	BP	circadian rhythm	1.01E-07	8	Up
GO:0006355	BP	regulation of transcription, DNA-templated	1.14E-07	20	Up
GO:0010017	BP	red or far-red light signaling pathway	9.22E-05	4	Up
GO:0032922	BP	circadian regulation of gene expression	2.31E-04	3	Up
GO:0010114	BP	response to red light	1.70E-05	4 / 2	Up/ Down
GO:0090351	BP	seedling development	1.93E-03	3	Up
GO:0000976	MF	transcription regulatory region sequence-specific DNA binding	1.93E-03	8	Up
GO:0003712	MF	transcription coregulator activity	4.65E-03	3	Up
GO:0010218	BP	response to far red light	7.44E-03	3	Up
GO:0005634	CC	nucleus	7.44E-03	43	Up
GO:0009507	CC	chloroplast	1.30E-11	25 / 34	Up/ Down
GO:0009649	BP	entrainment of circadian clock	1.39E-02	2 / 2	Up/ Down
GO:0005886	CC	plasma membrane	1.39E-02	19	Up
GO:0008324	MF	cation transmembrane transporter activity	1.47E-02	2	Up
GO:0031969	CC	chloroplast membrane	1.78E-02	2	Up
GO:0006865	BP	amino acid transport	2.11E-02	2	Up
GO:0009909	BP	regulation of flower development	2.37E-02	3	Up
GO:0009585	BP	red, far-red light phototransduction	2.55E-02	2	Up
GO:0003677	MF	DNA binding	2.75E-02	6	Up
GO:2000031	BP	regulation of salicylic acid mediated signaling pathway	2.86E-02	2	Up
GO:0010187	BP	negative regulation of seed germination	2.86E-02	2	Up
GO:0042651	CC	thylakoid membrane	3.20E-02	2	Up

GO:0043565	MF	sequence-specific DNA binding	3.37E-02	6	Up
GO:0080167	BP	response to karrikin	4.72E-02	3	Up
GO:0009658	BP	chloroplast organization	3.68E-08	8	Down
GO:0003729	MF	mRNA binding	5.83E-07	13	Down
GO:0003724	MF	RNA helicase activity	1.96E-05	5	Down
GO:0031425	BP	chloroplast RNA processing	8.19E-05	3	Down
GO:1901259	BP	chloroplast rRNA processing	1.52E-04	3	Down
GO:0000373	BP	Group II intron splicing	2.55E-04	3	Down
GO:0009570	CC	chloroplast stroma	2.55E-04	8	Down
GO:0009793	BP	embryo development ending in seed dormancy	4.49E-04	7	Down
GO:0009416	BP	response to light stimulus	8.09E-04	5	Down
GO:0009941	CC	chloroplast envelope	8.21E-04	7	Down
GO:0003723	MF	RNA binding	2.57E-03	6	Down
GO:0042793	BP	plastid transcription	3.57E-03	2	Down
GO:0042753	BP	positive regulation of circadian rhythm	3.61E-03	2	Down
GO:0009451	BP	RNA modification	5.67E-03	3	Down
GO:0009773	BP	photosynthetic electron transport in photosystem I	5.95E-03	2	Down
GO:0009535	CC	chloroplast thylakoid membrane	2.26E-02	4	Down
GO:0010027	BP	thylakoid membrane organization	2.47E-02	2	Down
GO:0009409	BP	response to cold	2.56E-02	4	Down
GO:0008270	MF	zinc ion binding	3.12E-02	5	Down
GO:0005618	CC	cell wall	3.12E-02	4	Down
GO:0009706	CC	chloroplast inner membrane	3.12E-02	2	Down
GO:0005829	CC	cytosol	3.27E-02	13	Down
GO:0009579	CC	thylakoid	3.36E-02	3	Down
GO:0009534	CC	chloroplast thylakoid	3.70E-02	3	Down
GO:0030246	MF	carbohydrate binding	4.88E-02	2	Down

Através das análises de enriquecimento, foi possível observar respostas muito diferentes entre as duas espécies no desenvolvimento nos substratos de estéril. *P. platycephala* teve uma resposta mais parecida com a obtida nos substratos naturais (tanto nos termos GO enriquecidos quanto nas vias metabólicas) do que *S. pulcherrimum*, indicando uma melhor aclimação de *P. platycephala* aos substratos de estéril. Para avaliar os mecanismos genéticos que fundamentam os diferentes desempenhos das espécies nos substratos de áreas degradadas, os genes ortólogos compartilhados foram utilizados para analisar o efeito da interação espécie-substrato em resposta ao cultivo em estéril vermelho e estéril amarelo. O teste não identificou genes ortólogos com um efeito significativo de interação espécie-substrato nas plantas cultivadas em estéril vermelho. No estéril amarelo, 51 genes ortólogos tiveram um efeito significativo de interação espécie-substrato (Tabela 13). A anotação funcional estava disponível para 22 destes genes (GO – 19 genes; KEGG – 8 genes). Nas análises de enriquecimento,

nenhuma via metabólica ou termo GO foram encontrados super-representados. Esses genes podem estar relacionados ao desempenho diferencial dessas espécies no substrato de estéril amarelo.

**Tabla 13:** Genes ortólogos diferencialmente regulados entre as espécies na resposta ao substrato de estéril amarelo.

Gene	Descrição	GO	KEGG
MTR_1g084190	xaa-Pro aminopeptidase	aminopeptidase activity mitochondrion peptidase activity	
MTR_3g086090	TPX2 (targeting protein for Xklp2) family protein		
MTR_5g043620	1-acyl-sn-glycerol-3-phosphate acyltransferase	1-acylglycerol-3-phosphate O-acyltransferase activity phosphatidic acid biosynthetic process	Metabolic pathways Biosynthesis of secondary metabolites Glycerophospholipid metabolism Glycerolipid metabolism
MTR_4g093570	beta-ureidopropionase-like protein	beta-ureidopropionase activity hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	Metabolic pathways Pyrimidine metabolism Pantothenate and CoA biosynthesis beta-Alanine metabolism
MTR_1g093350	auxin response factor		Plant hormone signal transduction
MTR_6g005390	sirohydrochlorin ferrochelatase		Metabolic pathways Biosynthesis of secondary metabolites Porphyrin and chlorophyll metabolism
MTR_8g024100	Serine/Threonine kinase family protein		
MTR_4g131750	DEAD-box ATP-dependent RNA helicase	mRNA splicing, via spliceosome RNA binding RNA helicase activity spliceosomal complex	
MTR_1g038430	DnaJ heat shock amino-terminal domain protein		
MTR_1g052730	exostosin family protein		
MTR_2g082930	hypothetical protein		
MTR_7g062360	aspartic proteinase nepenthesin	proteolysis	
MTR_4g068770	structural constituent of ribosome protein		
MTR_1g107495	hypothetical protein		
MTR_5g093440	DUF4408 domain protein		
MTR_7g118130	hypothetical protein		
MTR_5g088990	plastocyanin-like domain protein	anchored component of plasma membrane	
MTR_2g080290	transmembrane protein, putative		
MTR_8g465910	E3 ubiquitin-protein ligase RMA1H1-like protein	ubiquitin-dependent protein catabolic process ubiquitin-like protein conjugating enzyme binding ubiquitin protein ligase activity ER-associated misfolded protein catabolic process	Protein processing in endoplasmic reticulum

MTR_7g050980	pectinesterase/pectinesterase inhibitor	pectinesterase activity pectinesterase inhibitor activity	Metabolic pathways Pentose and glucuronate interconversions
MTR_1g116330	PRLI-interacting factor A		
MTR_8g080980	peptidyl-prolyl cis-trans isomerase FKBP62-like protein	protein peptidyl-prolyl isomerization peptidyl-prolyl cis-trans isomerase activity cytoplasm chaperone-mediated protein folding	
MTR_3g083680	PPR containing plant-like protein	chloroplast	
MTR_1g013110	PPPDE thiol peptidase family protein, putative	protein deubiquitination protein modification by small protein removal ubiquitinyl hydrolase activity	
MTR_8g036075	EF hand calcium-binding family protein	calcium ion binding	
MTR_1g073700	cyclopropane-fatty-acyl-phospholipid synthase	methyltransferase activity	
MTR_2g039910	calmodulin-binding family protein		
MTR_2g079550	glycoside hydrolase family 1 protein	Cyanoamino acid metabolism beta-glucosidase activity	Metabolic pathways Biosynthesis of secondary metabolites Starch and sucrose metabolism Phenylpropanoid biosynthesis
MTR_3g051220	ubiquitin-conjugating enzyme, putative		
MTR_2g014760	glutaredoxin-C1 protein		
MTR_4g115670	iron(III)-zinc(II) purple acid phosphatase, putative		
MTR_5g007040	zinc knuckle (CCHC-type) family protein, putative	Cdc73/Paf1 complex RNA polymerase II C-terminal domain phosphoserine binding	
MTR_0004s0230	hypothetical protein		
MTR_1g008220	CCT motif protein		
MTR_2g018820	PR containing plant-like protein		
MTR_3g048300	peptidase		
MTR_4g132890	acyl-protein thioesterase, putative	protein depalmitoylation cytoplasm palmitoyl-(protein) hydrolase activity carboxylic ester hydrolase activity	
MTR_2g084805	hypothetical protein		
MTR_7g022480	amidohydrolase family protein	cytoplasm secondary metabolic process	
MTR_2g008960	transmembrane protein, putative		
MTR_1g016430	calcineurin B-like protein 4-1		
MTR_5g085850	peptide/nitrate transporter	response to nitrate nitrate transmembrane transporter activity	
MTR_7g011850	plant/MXO21-9 protein		

MTR_2g072790	caffeic acid O-methyltransferase	O-methyltransferase activity S-adenosylmethionine-dependent methyltransferase activity aromatic compound biosynthetic process methylation	
MTR_1g107055	zinc finger-like protein		
MTR_4g115400	BSD domain protein		
MTR_7g090450	EF hand calcium-binding family protein		Plant-pathogen interaction
MTR_3g078240	NHL domain protein		
MTR_3g107860	hypothetical protein		
MTR_6g013720	zinc finger, C3HC4 type (RING finger) protein		
MTR_1g031650	calcium-dependent lipid-binding (CaLB domain) family protein		

## 5. Discussão

Neste trabalho apresentamos a montagem *de novo* do transcriptoma foliar para duas espécies não modelo de Fabaceae, *P. platycephala* e *S. pulcherrimum*, para as quais os dados de sequenciamento são limitados em bancos de dados públicos. Até o momento, apenas um genoma de cloroplasto para cada gênero (*Parkia javanica* - Número do bioprojeto: PRJNA390395 e *Stryphnodendron adstringens* - Número do bioprojeto: PRJNA573224) e 16 seqüências de DNA de *P. platycephala* e 18 de *S. pulcherrimum* estão disponíveis no banco de dados de Nucleotídeos do NCBI (Acessado em 02 de setembro de 2021). Portanto, este trabalho contribui de forma significativa para o conhecimento molecular desses dois gêneros. Os transcriptomas montados foram usados como referência para investigar a plasticidade da expressão gênica dessas espécies sob cultivo em dois substratos naturais e em dois tipos de substrato de estéril provenientes de áreas de mineração, com o intuito de avaliar a possibilidade de predição do sucesso de ocupação das áreas degradadas.

Uma etapa crucial para estudos genômicos de espécies não-modelo é a construção de uma referência de alta qualidade, uma vez que a qualidade e integridade da referência podem ter um impacto substancial na anotação e nas análises posteriores. Atualmente, diversas ferramentas estão disponíveis para a montagem de transcriptomas. Entretanto, nenhum *software* é capaz de fornecer o conjunto de transcritos uniformemente superior. Cada *software* tem seu método de reconstruir um transcriptoma *de novo* e pode ter um desempenho melhor em um subconjunto de loci específico ou identificar transcritos exclusivos em relação a outros (LU et al., 2013; GILBERT, 2013; VENTURINI et al., 2018). A combinação de montagens geradas por diferentes *softwares* e parâmetros pode levar à construção de uma montagem de transcriptoma mais abrangente (NAKASUGI et al., 2014; MAMROT et al., 2017; ROBERTS & ROALSON 2017; GILBERT, 2019). Dessa forma, uma abordagem de múltiplas montagens e parâmetros foi utilizada para gerar o melhor conjunto possível de transcritos. A qualidade das montagens das espécies foi avaliada, resultando em uma alta proporção de *reads* remapeadas nos transcriptomas e um alto número de ortólogos conservados completos recuperados.

Usando uma abordagem filogenética, foi possível fortalecer a confiança na precisão das montagens. Spillane e colaboradores (2020) mostraram que a qualidade da montagem do transcriptoma afeta as inferências filogenéticas de forma que montagens de alta qualidade levam a reconstruções filogenéticas mais precisas e contribuem para a consistência da filogenia já descrita. Os transcriptomas gerados foram utilizados para reconstruir a filogenia da subfamília Caesalpinioideae (LPWG, 2017) a partir de 331 ortogrupos de cópia única com

outros membros da subfamília disponíveis em bancos de dados públicos. A filogenia mostrou o posicionamento das duas espécies dentro do clado Mimosoidea e a relativa proximidade uma da outra. A topologia obtida apoia estudos realizados com a antiga subfamília Mimosoideae baseados em marcadores plastidiais (OLIVEIRA, 2015; SIMON et al., 2016; WANG, 2017a) e pelo primeiro esforço para reconstrução filogenética da família Fabaceae reclassificada usando dados plastidiais e nucleares (KOENEN et al., 2020). Esses resultados confirmam a precisão das montagens geradas nesse trabalho e mostram a proximidade evolutiva entre as espécies estudadas, o que dá suporte ao extrapolamento das nossas conclusões para clados próximos.

Os transcriptomas montados foram utilizados para avaliar as diferenças transcricionais de *P. platycephala* e *S. pulcherrimum* em resposta ao crescimento em substratos de canga, mata e dois resíduos de mineração. Embora o experimento tenha sido conduzido em casa de vegetação para controlar variáveis adicionais, todos os substratos utilizados foram coletados diretamente da área de ocorrência, mantendo suas características originais de campo (SILVA et al., 2018). O objetivo era compreender a resposta adaptativa geral aos diferentes substratos e não dissecar a contribuição de cada variável do solo individualmente. Estudos que investigam as respostas da expressão gênica a estímulos abióticos em condições naturais, com múltiplos estresses concomitantes, observam padrões transcriptômicos diferentes daqueles observados em ambientes altamente controlados (GUEVARA et al., 2012; SINHA et al., 2017; FAJARDO-CAVAZOS et al., 2018; CHO et al., 2018). Assim, mantendo os substratos originais, podemos nos aproximar dos ambientes naturais e fornecer uma melhor indicação das respostas transcriptômicas às condições da canga e das áreas de recuperação.

As plantas respondem a variações ambientais tais como temperatura, disponibilidade de água e nutrientes no solo e radiação. A resposta inicial das plantas a essas alterações normalmente se dá pela produção de metabólitos secundários relacionados a comunicação do que ocorre no ambiente visando a manutenção da homeostase na planta ou a defesa a estresses abióticos e bióticos (JAN et al., 2021). Nas análises de plasticidade da expressão gênica das plantas nos substratos naturais foram observadas alterações nas vias metabólicas relacionadas a biossíntese de metabólitos secundários em ambas as espécies. Vias relacionadas ao metabolismo primário também foram alteradas e podem estar relacionadas à produção dos metabólitos secundários (ISAH, 2019; JAN et al., 2021). Aproximadamente 52% das vias alteradas em *P. platycephala* e 75% das vias alteradas em *S. pulcherrimum* foram compartilhadas entre as espécies, a maioria relacionada ao metabolismo primário. As vias relacionadas à metabólitos secundários diferem nas duas espécies, indicando que, apesar do

número de vias compartilhadas, as espécies possuem estratégias diferentes para prosperar nos substratos naturais. As mudanças transcricionais também mostraram processos biológicos exclusivamente alterados para cada espécie.

Para seu crescimento e desenvolvimento, as plantas devem otimizar a absorção e alocação de carbono a partir da aquisição de recursos limitados, como água, luz e nutrientes. Quando os recursos estão disponíveis em quantidade ótima, os carboidratos produzidos por meio da fotossíntese são usados para maximizar a aptidão da planta que inclui o crescimento dos tecidos, reprodução, respiração celular, produção de metabólitos secundários e o armazenamento de carboidratos não estruturais (LONG et al., 2017). Os carboidratos não estruturais (CNEs), principalmente açúcares e amido, cumprem papéis funcionais distintos, incluindo transporte, metabolismo energético e osmorregulação, e fornecem substratos para a síntese de compostos de defesa ou para a troca com microrganismos envolvidos na aquisição de nutrientes (HARTMANN e TRUMBORE, 2016). Em locais onde a competição pela luz solar é alta e não há limitação na disponibilidade de nutrientes, as plantas favorecem o crescimento às custas do armazenamento de CNEs. Por outro lado, onde a disponibilidade de recursos é limitada, a alocação de recursos para o armazenamento de CNEs é favorecida (LONG et al., 2017). *P. platycephala* exibiu 13 genes alterados na comparação dos substratos naturais relacionados à biossíntese de fenilpropanoides, incluindo genes relacionados à biossíntese de lignina. A lignina é um polímero orgânico estrutural chave para o crescimento e desenvolvimento das plantas, conferindo rigidez à parede celular, fornecendo suporte estrutural e atuando como barreira contra patógenos. Também pode estar envolvida na nutrição mineral e na resposta das plantas a diversos estresses ambientais, como tolerância à seca, calor e metais pesados (DIAZ et al., 2001; GAO et al., 2012; BHARDWAJ et al., 2014; LIU et al., 2018), comumente observados em ambiente de canga. O aumento na biossíntese de lignina pode estar relacionado à alocação de recursos para o crescimento das plantas dessa espécie no substrato da canga. No substrato de mata, as plantas de ambas as espécies parecem alocar recursos para o armazenamento de CNEs. Esses resultados indicam a menor disponibilidade de nutrientes para as plantas no substrato de mata em comparação ao substrato de canga.

*P. platycephala* exibiu regulação negativa de genes relacionados à fotossíntese e fixação de carbono quando cultivada em substrato de canga. A regulação negativa de genes que codificam proteínas fotossintetizantes tem sido observada em estresses abióticos, como seca, sal, temperatura e metais pesados (BRESTIC et al., 2016, TAHMASEBI et al., 2019; COHEN e LEACH, 2019). A alta concentração de metais no tecido fotossintético pode levar a reduções na síntese de pigmentos fotossintéticos e danos à maquinaria fotossintética (PAUNOV et al.,

2018). Os mecanismos adotados pelas plantas neste estudo estão principalmente relacionados às consequências negativas na biossíntese da clorofila, na formação dos fotossistemas e nos mecanismos de transporte de elétrons. Genes relacionados à via de fixação de carbono também foram sub-expressos em *S. pulcherrimum* na canga, indicando aumento na expressão desses genes no substrato de mata. Apesar da modulação observada na expressão gênica, não foi observada alteração no acúmulo de biomassa ou no desempenho de crescimento entre os substratos naturais (SILVA et al., 2018). Dessa forma, as plantas das duas espécies favoreceram a alocação de recursos para o armazenamento de CNEs no substrato de mata sem prejuízo no crescimento.

O aumento na biossíntese de fenilpropanoides pode fazer parte de um dos mecanismos fisiológicos de *P. platycephala* envolvidos na tolerância ao metal, uma vez que elevada concentração de manganês (Mn) e ferro (Fe) foi observada nos tecidos foliares (SILVA et al., 2018). O envolvimento da lignina na tolerância ao metal foi relatado para *Phytolacca americana*, espécie conhecida acumuladora de Mn (GAO et al., 2012). O sequestro de Mn na parede celular da folha também contribui para a tolerância a Mn na cana-de-açúcar (YANG et al., 2019).

Os genes super-expressos da via de biossíntese dos fenilpropanoides incluem peroxidase e cinamil álcool desidrogenase, também envolvidos na regulação dos níveis de espécies reativas de oxigênio (ROS) (KIM et al., 2019). ROS podem alterar a integridade da estrutura celular e levar à desnaturação de proteínas funcionais e estruturais e à deterioração lipídica (KAPOOR et al., 2019). Outras vias alteradas de *P. platycephala* relacionadas à regulação de ROS foram peroxissomo e metabolismo de ascorbato e alderato. Dentre os GDEs foram identificados genes que codificam enzimas eliminadoras de ROS, como glutathione S-transferases, catalase (super-expressos), ascorbato peroxidase e superóxido dismutase (sub-expressos). Sarcosina oxidase, uma enzima que oxida sarcosina produzindo glicina, formaldeído e peróxido de hidrogênio (H<sub>2</sub>O<sub>2</sub>) nos peroxissomos, também foi encontrada diferencialmente expressa nas plantas cultivadas nos substratos naturais.

Termos GO relacionados à biossíntese de tiazol e a via do metabolismo de tiamina foram enriquecidos nos genes super-expressos das plantas de *P. platycephala*. A tiamina (vitamina B1) possui capacidade antioxidante (TUNC-OZDEMIR et al., 2009) e pode fornecer proteção contra danos oxidativos devido a altas concentrações de metal nessa espécie.

*S. pulcherrimum* exibiu genes regulados negativamente na biossíntese da estrutura dos terpenóides. Os terpenóides são a maior classe de metabólitos secundários em plantas e tem papéis importantes no alívio de estresses abióticos e bióticos (CHENG et al., 2007). Os genes

alterados nesta via codificam enzimas catalisadoras de reações que liberam pirofosfato (Geranilgeranil pirofosfato sintase, complexo de desidrodoliquil difosfato sintase, Solanesil-difosfato sintase 2, Geranilgeranil difosfato redutase). A regulação positiva na biossíntese de terpenóides observada em plantas cultivadas em substrato de mata pode estar relacionada a baixa disponibilidade de fósforo (P) no substrato (SILVA et al., 2018; Tabela 1). Além disso, as mudas de *S. pulcherrimum* cultivadas em substrato de mata apresentaram menor teor de P nas folhas (SILVA et al., 2018). Fósforo influencia a produção de terpenóides, uma vez que seus os precursores (IPP: difosfato de isopentenila e DMAPP: pirofosfato de dimetilalil) contêm ligações fosfato de alta energia e P é um componente chave de ATP e de NADPH, necessários para a síntese de terpenóides (ORMEÑO e FERNANDEZ, 2012; BUSTAMANTE et al., 2020). Assim, os terpenóides parecem atuar como fornecedores de fosfato em condições P-limitantes.

O substrato de mata utilizado nesse trabalho possuía menor quantidade de fósforo do que o substrato de canga (SILVA et al., 2018). Apesar de ser um nutriente limitante para o crescimento das plantas, nenhuma das duas espécies apresentaram sintomas físicos relacionados à deficiência de fósforo (SILVA et al., 2018). Entretanto, as análises de enriquecimento indicaram uma resposta de *S. pulcherrimum* a essa condição com 15 GDEs associados ao termo GO ‘resposta à deficiência de fosfato’. Vias metabólicas associadas ao metabolismo de lipídios e de galactose também foram enriquecidas nas análises e indicam um mecanismo da espécie para lidar com essa deficiência. Genes relacionados a síntese de monogalactosildiacilglicerol (MGDG), digalactosildiacilglicerol (DGDG) e sulfoquinovosildiacilglicerol (SQDG) foram encontrados sub-expressos nas análises, indicando super-expressão nas plantas cultivadas em substrato de mata. Esses compostos são galactolipídios (MGDG e GDGD) e sulfolipídios (SQDG) que compõe a maioria dos lipídios de membrana do cloroplasto, tornando a organela minimamente dependente de fosfato (COOK et al., 2021). Ainda assim, os fosfolipídios compõem até 15% das membranas do cloroplasto (COOK et al., 2021). Durante exposição à privação de fósforo, as plantas passam a realocar fosfato através da troca dos lipídios de membrana do cloroplasto (YU et al., 2002; KOBAYASHI et al., 2009). A biossíntese de galactolipídios e sulfolipídios é aumentada, ocorrendo a substituição dos fosfolipídios e liberando fosfato para manutenção dos níveis de ácidos nucleicos e a atividade metabólica (YU et al., 2002; KOBAYASHI et al., 2009; COOK et al., 2021). Dessa forma, *S. pulcherrimum* parece ter prosperado no substrato de mata apesar da deficiência de fosfato através da troca dos lipídios de membrana do cloroplasto.

As plantas respondem de maneiras diferentes a um mesmo ambiente, entretanto condições ambientais adversas limitam a gama de estratégias fisiológicas ou ecológicas e levam

à convergência de características. Essa convergência pode ou não ser resultado da expressão de um mesmo conjunto de genes. Dessa forma, a possibilidade de as espécies estudadas exibirem estratégias moleculares comuns subjacentes às adaptações aos substratos foi avaliada. A análise comparativa da expressão gênica dos ortólogos revelou que os padrões de expressão diferiam mais entre as espécies do que entre as condições de crescimento, indicando que o padrão geral da expressão gênica é específico do organismo (Figura 6c-d). Mesmo assim quase 300 pares de ortólogos foram identificados respondendo de maneira similar nas duas espécies na comparação entre o crescimento no substrato de canga e de mata. Esses transcritos codificam proteínas envolvidas no ritmo circadiano. A via do ritmo circadiano e os termos GO também foram enriquecidos nas análises espécie-específicas. O ritmo circadiano é sincronizado por mudanças nos estímulos luminosos e na temperatura (DODD et al., 2005; YERUSHALMI & GREEN, 2009) e permite que as plantas antecipem mudanças diárias e sazonais no ambiente que são essenciais para regular seu crescimento e necessárias para um desempenho bem-sucedido (CALDEIRA et al., 2014). O mecanismo do relógio circadiano nas plantas consiste em uma rede de genes com vários loops de feedback transcricional (MCCLUNG, 2019). Em *A. thaliana*, um par de fatores de transcrição intimamente relacionados, *Late Elongated Hypocotyl* (LHY) e *Circadian Clock Associated* (CCA1), expressos ao amanhecer reprimem a expressão de outros componentes do relógio expressos a noite, dentre eles o conjunto de *pseudo-response regulators* (PRR), *GIGANTEA* (GI) e “*Evening Complex*” formado por *LUX ARRHYTHMO* (LUX), *EARLY FLOWERING* (ELF3) e ELF4. Os PRRs são sequencialmente expressos e reprimem a transcrição de CCA1 e LHY, bem como sua própria transcrição. À tarde, a ativação transcricional é mediada pelos genes *NIGHT LIGHT-INDUCIBLE AND CLOCK-REGULATED* (LNKs), que promovem a transcrição de PRR9, PRR5, TOC1 (ou PRR1), GI, LUX e ELF4. A ativação de ELF4 também é mediada pelo *ELONGATED HYPOCOTYL 5* (HY5). Vários estudos demonstraram que o relógio circadiano contribui para a capacidade das plantas de tolerar um amplo espectro de condições estressantes e se aclimatar a elas, incluindo deficiência de ferro, estresse alcalino e estresse por seca ou salinidade (GRUNDY et al., 2015; LI et al., 2019). Neste estudo, o crescimento no substrato de canga regulou negativamente a expressão de LNKs (1, 2 e 3), LHY e HY5, e regulou positivamente genes que são expressos a noite (PCL1, ELF4, PRR5) e *COLD-REGULATED GENE* (CORs 27 e 28). O perfil de expressão gênica do relógio circadiano observado aqui parece estar relacionado à regulação negativa das funções do cloroplasto. A regulação circadiana está integrada com a fotossíntese, o metabolismo de fixação de carbono e seus produtos metabólicos (HAYDON et al., 2013; DODD et al., 2015). CCA1, homólogo de LHY em *Arabidopsis*, é super-expresso em resposta

a adição de açúcar. Por outro lado, a expressão de *PRR7* é reprimida (HAYDON et al., 2013). De fato, as vias relacionadas ao metabolismo do carbono, como a fixação do carbono em organismos fotossintéticos, glicólise e vias da pentose fosfato, foram enriquecidas nos genes regulados negativamente em ambas as espécies nas análises espécie-específicas. Portanto, a composição do substrato afetou a expressão dos genes do ritmo circadiano nas folhas, regulando o metabolismo de fixação do carbono. Isso pode fornecer adaptações para otimizar o desempenho das plantas em ambientes com diferentes condições nutricionais. O relógio circadiano tem um papel na regulação da homeostase de micronutrientes em plantas, incluindo aquisição e transporte para a parte aérea (PEREA-GARCÍA et al., 2010; WANG et al., 2011; CHEN et al., 2013). O aumento na expressão dos genes do ritmo circadiano que são expressos de manhã nas plantas cultivadas no substrato de mata pode estar relacionado a baixa disponibilidade de nutrientes nesse substrato. Wang e colaboradores (2011) mostraram que o transportador de fosfato *PHT4;1* (*PHOSPHATE TRANSPORTER 4;1*) é regulado pelo ritmo circadiano e tem seu pico de expressão de manhã, consistente com o aumento da demanda por fotossíntese durante o dia. Dessa forma, a super-expressão dos genes do ritmo circadiano no substrato de mata pode levar ao aumento dos transportadores desse nutriente.

Outra possibilidade é que as plantas cultivadas no substrato de canga tenham alterado a fase do relógio circadiano, avançando a expressão de genes que expressam a noite, já que as amostras biológicas foram coletadas pela manhã. Um avanço semelhante na fase do relógio circadiano também foi observado na cevada sob estresse osmótico (HABTE et al., 2014). Chen e colaboradores (2013) mostraram que a deficiência de Fe prolongou o período do ritmo circadiano em *Arabidopsis*. Tanto *P. platycephala* quanto *S. pulcherrimum* no substrato da canga exibiram alta concentração de Fe nas folhas (SILVA et al., 2018), o que pode estar relacionado a um período circadiano encurtado e explicar a expressão precoce dos genes noturnos. Observar a expressão dos genes do ritmo circadiano em vários momentos durante o dia pode elucidar se o substrato da canga encurta o período circadiano. De maneira geral, muitos genes envolvidos na resposta ao estresse abiótico estão sob o controle do ritmo circadiano, mesmo para condições ambientais que são constantes de maneira diurna, como seca e salinidade (CALDEIRA et al., 2014; GRUNDY et al., 2015). Nossos resultados sugerem que as plantas adaptadas aos ambientes de canga e floresta modulam o ritmo circadiano de uma maneira dependente da composição do substrato que pode ajudá-las a prosperar nessa gama de condições. O relógio circadiano é conservado entre as espécies vivas, pois controla os processos metabólicos gerais e garante a aclimação das plantas ao seu ambiente. Uma possibilidade interessante é investigar se as plantas endêmicas da canga apresentam menor plasticidade na

expressão de genes do ritmo circadiano, limitando sua capacidade de ocupar ambientes variados. A modificação dos genes do relógio circadiano pode aumentar o crescimento e a produtividade de culturas (LI & LAM, 2020; STEED et al., 2021). Da mesma forma, a modulação dos genes do relógio circadiano pode melhorar a adaptação local ao ambiente pela percepção do estado nutricional.

Após a avaliação da plasticidade da expressão gênica nos substratos naturais, a resposta das espécies ao crescimento em substratos desafiadores de estéril de mina foi testada. O perfil de expressão gênica das espécies cultivadas em estéril vermelho foi parecido com o crescimento em canga, com poucos GDEs. *P. platycephala* não apresentou nenhum termo GO enriquecido e apenas a via do metabolismo de glicerofosfolípidios foi alterada no crescimento em estéril vermelho quando comparado com os substratos naturais. Em estéril amarelo, os mesmos termos GO e a maior parte das vias metabólicas foram enriquecidas em comum com a comparação dos substratos naturais. *S. pulcherrimum* teve um maior número de vias metabólicas alteradas com relação ao crescimento nos substratos naturais em comparação ao cultivo em estéril amarelo. Muitos termos GO foram enriquecidos em estéril vermelho, apesar do baixo número de GDEs, a maior parte relacionada à deficiência de fósforo.

A resposta de *S. pulcherrimum* com relação a deficiência de fósforo em estéril vermelho foi similar à encontrada no crescimento em substrato de mata, com a super-expressão de genes relacionados aos lipídios de membrana não fosfatados. Dessa forma, apesar do substrato de estéril vermelho ter mais fósforo do que o de mata (Tabela 1; SILVA et al., 2018), ainda foi um fator limitante para *S. pulcherrimum*. A maior acidez do solo pode ter limitado a disponibilidade desse nutriente para as plantas, uma vez que em solos mais ácidos, o fósforo forma compostos com Fe, cálcio e alumínio que precipitam (VANCE et al., 2003), justificando a variação da expressão gênica em estéril vermelho para *S. pulcherrimum*. A espécie também apresentou menor incidência de associações com bactérias fixadoras em estéril vermelho do que em canga e em mata (SILVA et al., 2018).

Um grande número de genes relacionados à percepção de estímulos bióticos e à resposta do sistema imune foram encontrados sub-expressos nas plantas de *S. pulcherrimum* cultivadas em estéril amarelo (processos biológicos enriquecidos: resposta a estímulo biótico, resposta a estímulo biótico externo, resposta a outros organismos, resposta de defesa a outro organismo, resposta a fungos, resposta de defesa à fungos, resposta à bactéria, resposta imune, processo do sistema imune, resposta imune inata). A via metabólica interação planta-patógeno também foi encontrada enriquecida nos genes sub-expressos. Esses resultados poderiam sugerir que as folhas das plantas cultivadas na canga teriam sido infectadas por fungos ou bactérias, o que

justificaria a super-expressão desses genes, ainda que nenhum sintoma tivesse sido observado nas folhas no momento da coleta. Entretanto, todas as outras comparações (mata e estéril vermelho) foram realizadas contra a condição da canga e em nenhuma delas esses genes foram alterados. Dessa forma, esses resultados podem estar relacionados à resposta das plantas às associações com microrganismos nas raízes. Silva e colaboradores (2018) observaram que *S. pulcherrimum* produziu estruturas nodulares, que indicam interações simbióticas com bactérias fixadoras de nitrogênio, no crescimento em todos os substratos, com exceção de estéril amarelo. Microrganismos simbióticos, como fungos micorrízicos e rizobactérias podem induzir resistência sistêmica nas plantas, aumentando a expressão de genes relacionados à defesa em tecidos não infectados (ZAMIOUDIS e PIETERSE, 2012; NISHAD et al., 2020). A associação com rizobactérias parece ter induzido a resposta imune em *S. pulcherrimum* através do aumento na expressão de WRKY e ERF (fator de resposta à etileno), fatores de transcrição que regulam genes envolvidos na defesa e potencializam a resposta imune das plantas após uma infecção (DEY et al., 2014). As plantas em estéril amarelo não apresentaram indução da resposta imune sistêmica, evidenciado pela regulação negativa dos genes em comparação com plantas em canga, deixando-as mais suscetíveis ao ataque por patógenos.

A falta de associação com organismos fixadores de nitrogênio em estéril amarelo foi atribuída a menor quantidade de matéria orgânica e diversidade de microrganismos fixadores nesse substrato, especialmente aqueles exigidos por espécies especialistas (SILVA et al., 2018). *Chloroleucon acacioides* apresentou nódulos nos quatro substratos testados (SILVA et al., 2018) e foi considerada uma espécie generalista, apresentando associação com uma maior diversidade de microrganismos presentes em todos os substratos. Entretanto, a disponibilidade de nutrientes no solo e a necessidade específica de cada espécie por nutrientes também pode interferir na formação dessas associações (ZHRAN, 1999). A deficiência de fósforo está associada a diminuição da formação e desenvolvimento de nódulos (HERNÁNDEZ et al., 2009, ISIDRA-ARELLANO et al., 2018). A maior exigência de *S. pulcherrimum* por fósforo associada a menor disponibilidade desse nutriente em estéril amarelo pode ter impactado a nodulação dessa espécie nesse substrato. Dessa forma, a suplementação de P ou a aplicação de microrganismos solubilizadores de P no substrato de estéril amarelo pode melhorar a associação de *S. pulcherrimum* com bactérias fixadoras de nitrogênio.

Os ajustes metabólicos em resposta a condições desfavoráveis são dinâmicos e multifacetados e dependem da espécie e do tipo e da intensidade do estresse (KRASENSKY & JONAK, 2012). *S. pulcherrimum* apresenta estratégias para compensar a limitação de fósforo, não alterando a taxa de crescimento e ganho de biomassa nos substratos de mata e estéril

vermelho (SILVA et al., 2018). No substrato de estéril amarelo, a taxa de crescimento e o ganho de biomassa das plantas de *S. pulcherrimum* foram menores, indicando maior nível de estresse. Várias características desse substrato podem estar relacionadas a esse resultado, como a falta de associação com bactérias fixadoras de nitrogênio; maior deficiência de fósforo; altos níveis de enxofre e a compactação do solo, que podem limitar o aporte de nutrientes. Apesar da alocação de recursos para crescimento e desenvolvimento ter sido afetada, *S. pulcherrimum* apresentou ajustes no metabolismo para compensação do estresse em estéril amarelo e pode ser recomendada para recuperação de ambos os tipos de estéril. Porém, a exposição prolongada à falta de fósforo pode levar a uma perturbação irreversível no sistema. Dessa forma, os resultados observados nesse trabalho indicam a necessidade de suplementação de fósforo para melhor estabelecimento dessa espécie. Através das análises foi possível prever a necessidade nutricional das plantas, antes do aparecimento de sintomas, tornando possível a indicação de remediação do substrato a tempo de recuperação. Mesmo nos substratos em que não houve diferença na taxa de crescimento e no acúmulo de biomassa, foi possível observar que há uma limitação de nutrientes para o desenvolvimento das plantas dessa espécie e identificar qual é esse nutriente.

*P. platycephala* exibiu maior tolerância à limitação de nutrientes em todos os substratos (SILVA et al., 2018). Termos GO relacionados a deficiência de fósforo não foram encontrados super-representados nas análises para essa espécie. A necessidade de plantas não fixadoras, como *P. platycephala*, por P foi descrita como menor do que a das plantas fixadoras de nitrogênio, principalmente como consequência da grande demanda de P para os processos de fixação biológica (GUO et al., 2017). Essa espécie parece ter menor necessidade de fósforo para manutenção do metabolismo e homeostase do que *S. pulcherrimum*, sendo mais resistente a substratos com limitação desse nutriente. *P. platycephala* apresentou variações na expressão gênica consistentes com a adaptação às condições dos substratos. Esses resultados indicam o provável sucesso no estabelecimento dessa espécie e aclimatação às áreas de recuperação. O uso de *P. platycephala* poderia facilitar o processo de reabilitação uma vez que ela é menos dependente do manejo do substrato (ver GIANINI et al., 2016).

267 ODEsRS foram encontrados nas espécies em estéril amarelo. Apesar da análise de enriquecimento das vias metabólicas não ter retornado resultados significativos, os termos GO enriquecidos mostram que esses ortólogos também estão associados ao ritmo circadiano e a estímulos luminosos. O perfil de expressão dos ODEsRS em estéril amarelo foi parecido com os ODEsRS em mata, com alteração dos mesmos genes do ritmo circadiano (LNKs, LHY, UNE10 – Super-expressos; COR27, ELF4, PRR5, PCL1 – sub-expressos com relação ao

substrato de canga). Dessa forma, as plantas mostraram uma resposta circadiana no desenvolvimento nos diferentes substratos que pode estar relacionada com o favorecimento do aporte de nutrientes do solo.

Nossos resultados demonstram o poder da plasticidade da expressão gênica para promover respostas biológicas adaptativas às diferentes condições de crescimento das plantas. A determinação da base genômica para mudanças nas características fisiológicas e comportamentais associadas à adaptação pode melhorar a taxa de sucesso de reintrodução das espécies e conseqüentemente a recuperação das áreas. Nesse trabalho foi possível prever o sucesso de estabelecimento das plantas nos novos ambientes e ajudar na indicação das espécies para reabilitação. A mesma abordagem pode ajudar na escolha de populações-fonte para recuperação de áreas. Diferentes populações da mesma espécie, adaptadas a diferentes locais podem apresentar tolerâncias diferentes aos substratos das áreas em recuperação.

A falta de conhecimento sobre as espécies nativas e as diferenças entre as populações dificulta a escolha das espécies para revegetação e recuperação das áreas, bem como as estratégias para conservação. O conhecimento dos mecanismos fisiológicos melhora o entendimento das necessidades das espécies e a diagnose das ações necessárias para o sucesso de reintrodução das espécies nas áreas-alvo. O conhecimento da plasticidade da expressão gênica natural das espécies ajuda na interpretação da resposta fisiológica das plantas com relação a mudanças no ambiente e na identificação de fatores ambientais limitantes que podem levar ao prejuízo no estabelecimento das espécies.

## 6. Conclusões

O presente estudo apresenta a plasticidade da expressão gênica de duas espécies próximas de Fabaceae capazes de habitar ambientes de canga e de floresta. A partir desse conhecimento, também foi apresentada a predição da resposta fisiológica das espécies a substratos de áreas mineradas, cuja recuperação é mandatória. A alta qualidade dos transcriptomas montados *de novo* gerados para ambas as espécies permitiram a identificação de GDEs implicados em diversas vias metabólicas. A composição do substrato modula a expressão gênica foliar em *P. platycephala* e *S. pulcherrimum*. Ambas as espécies exibiram mudanças nas principais vias metabólicas, como a biossíntese de metabólitos secundários e o metabolismo do carbono, indicando alocação de recursos para armazenamento de carboidratos não estruturais nos substratos de mata e estéril amarelo. Nos substratos de canga e estéril vermelho, *P. platycephala* apresentou aumento na expressão de genes relacionados a biossíntese dos fenilpropanóides, relacionada a alocação de recursos para o crescimento da planta. *S. pulcherrimum* apresenta resposta à limitação de fósforo nos substratos de mata e de resíduos de mineração, provavelmente associada a necessidade desse nutriente para fixação biológica de nitrogênio. A espécie apresenta um conjunto de estratégias para compensar essa limitação de forma eficiente como a troca dos fosfolipídios de membrana e a liberação de fosfato na via dos terpenóides. Em contraste, *P. platycephala* parece apresentar menor necessidade de fósforo para manutenção do metabolismo. A maior disponibilidade de fósforo no substrato da canga possibilitou a super-expressão de genes relacionados a biossíntese de tiamina nessa espécie, que pode estar relacionada a proteção contra ROS provenientes do acúmulo de metal nas folhas. Os padrões observados sugerem que cada uma das espécies estudadas emprega seu próprio catálogo de genes na resposta aos diferentes substratos de cultivo. Nossas análises também revelaram padrões de plasticidade na expressão gênica compartilhados entre as espécies aos substratos. A modulação de genes do ritmo circadiano foi um mecanismo comum relacionado a resposta a limitação de nutrientes nos substratos de mata e estéril amarelo.

O presente estudo fornece informações valiosas sobre a base genética da resposta à limitação de nutrientes no solo em *P. platycephala* e *S. pulcherrimum*. A compreensão desses mecanismos pode ter implicações importantes para a conservação das espécies e para a recuperação de áreas degradadas. Os resultados obtidos podem servir de referência para futuros

trabalhos sobre a base genética de estresses abióticos nessas duas espécies e em outras espécies de leguminosas.

## 7. Referências

- Almeida ROPO, Sanchez LE. Revegetação de áreas de mineração: 747 critérios de monitoramento e avaliação do desempenho. *Revista Árvore* 2005;29:47-54
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 1990;215(3):403–410
- Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Assefa AT, Vandesompele J, Thas O. On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments. *BMC Genomics* 2020; 21:312. doi:10.1186/s12864-020-6721-y.
- Babiychuk E, Kushnir S, Vasconcelos S, Dias MC, Carvalho-Filho N, Nunes GL, et al. Natural history of the narrow endemics *Ipomoea cavalcantei* and *I. marabaensis* from Amazon Canga savannahs. *Sci. Rep.* 2017;7:7493.
- Babu CVS, Gassmann M. Assessing Integrity of Plant RNA with the Agilent 2100 Bioanalyzer System. *Application Notes. Genomics.* 20016;5990-8850EN.
- Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 2012;19(5):455–477; doi:10.1089/cmb.2012.0021.
- Barr S, Jonas JL, Paschke MW. Optimizing seed mixture diversity and seeding rates for grassland restoration. *Restor. Ecol.* 2017;25:396–404.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R.Stat.Soc.Ser.B* 1995;57:289–300.
- Benning C. MAFFT - a multiple sequence alignment program. *Annual Review of Cell and Developmental Biology* 2009;25:71–91.
- Bhardwaj R, Handa N, Sharma R, Kaur H, Kohli S, Kumar V, Kaur P. Lignins and abiotic stress: An overview. In: Ahmad P, Wani M, editors. *Physiological Mechanisms and Adaptation Strategies in Plants Under Changing Environment*. Springer: New York. 2014; doi:10.1007/978-1-4614-8591-9\_10.
- Bolger AM, Lohse M, Usadel B Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
- Brestic M, Zivcak M, Kunderlikova K, Allakhverdiev SI. High temperature specifically affects the photoprotective responses of chlorophyll b-deficient wheat mutant lines. *Photosynth. Res.* 2016;130:251–266.

- Bryant DM, Johnson K, DiTommaso T, et al. (2017) A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell reports* 2017;18(3):762-776; doi:10.1016/j.celrep.2016.12.063.
- Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*, 2016.
- Bustamante MÁ, Michelozzi M, Barra Caracciolo A, et al. Effects of Soil Fertilization on Terpenoids and Other Carbon-Based Secondary Metabolites in *Rosmarinus officinalis* Plants: A Comparative Study. *Plants (Basel)*. 2020;9(7):830; doi:10.3390/plants9070830
- Caldeira CF, Jeanguenin L, Chaumont F, Tardieu F. Circadian rhythms of hydraulic conductance and growth are enhanced by drought and improve plant performance. *Nat Commun*. 2014;5:5365; doi:10.1038/ncomms6365.
- Carmo FF, Carmo, FF, Campos IC, Jacobi CM. Cangas: ilhas de ferro estratégicas para a conservação. *Ciência Hoje*, 2012;295: 48-53.
- Carvalho CS, Lanes ECM, Silva AR, Caldeira CF, Carvalho-Filho N, Gastauer M, Imperatriz-Fonseca VL, Nascimento Júnior W, Oliveira G, Siqueira JO, Viana PL, Jaffé R. Habitat Loss Does Not Always Entail Negative Genetic Consequences. *Front Genet*. 2019;13(10):1011. doi:10.3389/fgene.2019.01101.
- Cicatelli A, Lingua G, Todeschini V, Biondi S, Torrigiani P, Castiglione S. Arbuscular mycorrhizal fungi restore normal growth in a white poplar clone grown on heavy metal-contaminated soil, and this is associated with upregulation of foliar metallothionein and polyamine biosynthetic gene expression *Ann. Bot*. 2010;106, 791-802.
- Chen YY, Wang Y, Shin LJ, Wu JF, Shanmugam V, Tsednee M, Lo JC, Chen CC, Wu SH, Yeh KC. Iron is involved in the maintenance of circadian period length in *Arabidopsis*. *Plant physiology*, 2013;161(3):1409–1420; doi:10.1104/pp.112.212068.
- Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* 2016, 5:1438; doi: 10.12688/f1000research.8987.2
- Cheng A, Lou Y, Mao Y, Lu S, Wang L, Chen X. Plant terpenoids: Biosynthesis and ecological functions. *J. Integr. Plant. Biol*. 2007, 49:179–186; doi:10.1111/j.1744-7909.2007.00395.x.
- Cho SM, Lee H, Jo H, Lee H, Kang Y, Park H, Lee J. Comparative transcriptome analysis of field- and chamber-grown samples of *Colobanthus quitensis* (Kunth) Bartl, an Antarctic flowering plant. *Sci Rep* 2018;8:11049; doi:10.1038/s41598-018-29335-4.
- Chikhi R, Medvedev P. Informed and Automated k-Mer Size Selection for Genome Assembly. *Bioinformatics* 2014;30(1):31-7.
- Cohen SP, Leach JE. Abiotic and biotic stresses induce a core transcriptome response in rice. *Sci. Rep*. 2019;9:6273; doi:10.1038/s41598-019-42731-8.

Cook R, Lupette J, Benning C. The Role of Chloroplast Membrane Lipid Metabolism in Plant Environmental Responses. *Cells* 2021, 10:706; <https://doi.org/10.3390/cells10030706>

Cooke JA, Johnson MS. Ecological restoration of land with particular reference to the monitoring of metals and industrial minerals: a review of theory and practice. *Environ. Rev.* 2002;10:41–71. doi: 10.1139/a01-014

Dey S, Wenig M, Langen G, et al. Bacteria-triggered systemic immunity in barley is associated with WRKY and ETHYLENE RESPONSIVE FACTORS but not with salicylic acid. *Plant Physiol.* 2014; 166(4):2133-2151. doi:10.1104/pp.114.249276

Diaz J, Bernal A, Pomar F, Merino F. Induction of shikimate dehydrogenase and peroxidase in pepper (*Capsicum annuum* L.) seedlings in response to copper stress and its relation to lignification. *Plant Sci.* 2001;161:179–188.

DNPM – Departamento Nacional de Produção Mineral. 2017 Sumário Mineral - Ferro. Coordenadores: Jesus, C.A.G. & Joaquim, L.G. Disponível em: [http://www.anm.gov.br/dnpm/publicacoes/serie-estatisticas-e-economia-mineral/sumario-mineral/ferro\\_sm\\_2017](http://www.anm.gov.br/dnpm/publicacoes/serie-estatisticas-e-economia-mineral/sumario-mineral/ferro_sm_2017)

Dodd AN, Salathia N, Hall A, Kévei E, Tóth R, Nagy F, Hibberd JM, Millar AJ, Webb AA. Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science* 2005;309(5734):630-633; doi:10.1126/science.1115581.

Dodd AN, Belbin FE, Frank A, Webb AA. Interactions between circadian clocks and photosynthesis for the temporal and spatial coordination of metabolism. *Front Plant Sci.* 2015;6:245. doi:10.3389/fpls.2015.00245

Dorr JVN. Physiographic, stratigraphic and structural development of Quadrilátero Ferrífero, Minas Gerais, Brasil. Prof. Paper 641-A. USGS; Washington: 110. 1969

Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology.* 2015;16:157.

Emms DM, Kelly S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution.* 2017;34(12):3267–3278. doi:10.1093/molbev/msx259.

Emms D, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 2019;20:238.

Fajardo-Cavazos P, Morrison MD, Miller KM, Schuerger AC, Nicholson WL. Transcriptomic responses of *Serratia liquefaciens* cells grown under simulated Martian conditions of low temperature, low pressure, and CO<sub>2</sub>-enriched anoxic atmosphere. *Sci Rep.* 2018;8(1):14938; doi:10.1038/s41598-018-33140-4.

Faria SM, Moraes LFD, Lima HC, Ribeiro RD, Mattos CMJ, Rodrigues TR, Castilho AF, Canosa GA, Silva MAP. Composição florística de leguminosas com potencial para fixação biológica de nitrogênio em áreas de vegetação de canga (savana metalófito) do entorno do complexo minerador de Carajás. Seropédica: EMBRAPACNPAB, Comunicado Técnico 2011;140: 1-20.

Forzza RC, Baumgratz JFA, Bicudo CEM, Carvalho Jr AA, Costa A, Costa DP. Catálogo de plantas e fungos do Brasil. Rio de Janeiro: Andrea Jakobsson Estúdio - Instituto de Pesquisa Jardim Botânico do Rio de Janeiro, 2010; 828 p. Vol. 2. ISBN 978-85- 8874-243-7. Available from SciELO Books <<http://books.scielo.org>>.

Gao L, Peng K, Chen Y, Wang G, Shen Z. Roles of apoplastic peroxidases, laccases, and lignification in the manganese tolerance of hyperaccumulator *Phytolacca americana*. *Acta Physiol. Plant* 2012;34:151–159.

Garg R, Patel RK, Tyagi AK, Jain M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* 2011;18(1):53-63. doi:10.1093/dnares/dsq028

Gastauer M, Silva JR, Caldeira CF, Ramos SJ, Souza Filho PWM, Furtini Neto AE, Siqueira JO. Mine land rehabilitation: Modern ecological approaches for more sustainable mining. *J. Clean. Prod.* 2018;172:1409–1422.

Gastauer M, Sarmiento PSM, Santos VCA, Caldeira CF, Ramos SJ, Teodoro GS, Siqueira JO. Vegetative functional traits guide plant species selection for initial mineland rehabilitation. *Ecological Engineering.* 2020;148; doi:10.1016/j.ecoleng.2020.105763.

Ghalambor CK, McKay JK, Carroll SP, Reznick DN. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology* 2007;21:394–407.

Giannini TC, Giuliatti AM, Harley RM, et al. Selecting plant species for practical restoration of degraded lands using a multiple-trait approach. *Austral Ecology.* 2016; doi:10.1111/aec.12470.

Gilbert D. Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame. 2013. <http://arthropods.eugenes.org/EvidentialGene/about/EvigeneRNA2013poster.pdf> doi:10.7490/fl1000research.1112594.1.

Gilbert D. Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? Preprint at bioRxiv 829184. 2019; doi:10.1101/829184.

Grabherr MG., Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 2011;29(7):644–652; doi:10.1038/nbt.1883.

Grundy J, Stoker C, Carré IA. Circadian regulation of abiotic stress tolerance in plants. *Front Plant Sci.* 2015;6:648; doi:10.3389/fpls.2015.00648.

Guevara DR, Champigny MJ, Tattersall A, et al. Transcriptomic and metabolomic analysis of Yukon *Thellungiella* plants grown in cabinets and their natural habitat show phenotypic plasticity. *BMC plant biology* 2012;12:175; doi:10.1186/1471-2229-12-175.

- Guo Y, Yang X, Schöb C, Jiang Y, Tang Z. Legume Shrubs Are More Nitrogen-Homeostatic than Non-legume Shrubs. *Front Plant Sci.* 2017;8:1662; doi:10.3389/fpls.2017.01662
- Habte E, Müller LM, Shtaya M, Davis SJ, von Korff M. Osmotic stress at the barley root affects expression of circadian clock genes in the shoot. *Plant Cell Environ.* 2014;37(6):1321-7. doi:10.1111/pce.12242.
- Hartmann H, Trumbore S. Understanding the roles of nonstructural carbohydrates in forest trees - from what we can measure to what we want to know. *New Phytol.* 2016;211(2):386-403; doi: 10.1111/nph.13955.
- Haydon MJ, Bell LJ, Webb AA. Interactions between plant circadian clocks and solute transport. *J. Exp. Bot.* 2011;62:2333–2348.
- He X, Johansson ML, Heath DD. Role of genomics and transcriptomics in selection of reintroduction source populations. *Conserv Biol.* 2016;30(5):1010-8; doi:10.1111/cobi.12674.
- Hernández G, Valdés-López O, Ramírez M, Goffard N, Weiller G, Aparicio-Fabre R, Fuentes SI et al. Global changes in the transcript and metabolic profiles during symbiotic nitrogen fixation in phosphorus-stressed common bean plants. *Plant Physiol.* 2009, 151:1221–1238.
- Hsieh PH, Oyang YJ, Chen CY. Effect of de novo transcriptome assembly on transcript quantification. *Sci. Rep.* 2019;9:8304; doi:10.1038/s41598-019-44499-3.
- Isah T. Stress and defense responses in plant secondary metabolites production. *Biol. Res.* 2019;52(1):39; doi:10.1186/s40659-019-0246-3.
- Isidra-Arellano MC, Reyero-Saavedra MDR, Sánchez-Correa MDS. *et al.* Phosphate deficiency negatively affects early steps of the symbiosis between common bean and rhizobia. *Genes*, 2018;9:E498.
- Jacobi CM, Carmo FF, Vincent RC, Stehmann JR. Plant communities on the ironstone outcrops - a diverse and endangered Brazilian ecosystem. *Biodivers. Conserv.* 2007;16:2185–2200; doi:10.1007/s10531-007-9156-8.
- Jacobi CM, Carmo FF, Campos IC. Soaring extinction threats to endemic plants in Brazilian metal-rich regions. *Ambio* 2011;40:540–543.
- Jan R, Asaf S, Numan M, Lubna, Kim K-M. Plant Secondary Metabolite Biosynthesis and Transcriptional Regulation in Response to Biotic and Abiotic Stress Conditions. *Agronomy.* 2021; 11(5):968; doi:0.3390/agronomy11050968
- Kapoor D, Singh S, Kumar V, Romero R, Prasad R, Singh J. Antioxidant enzymes regulation in plants in reference to reactive oxygen species (ROS) and reactive nitrogen species (RNS). *Plant Gene* 2019;19; doi:10.1016/j.plgene.2019.100182.
- Kim YH, Huh GH. Overexpression of cinnamyl alcohol dehydrogenase gene from sweetpotato enhances oxidative stress tolerance in transgenic *Arabidopsis*. *In Vitro Cell.Dev.Biol.-Plant* 2019;55:172–179; doi:10.1007/s11627-018-09951-5.

- Kobayashi K, Awai K, Nakamura M, Nagatani A, Masuda T, Ohta H. Type-B monogalactosyldiacylglycerol synthases are involved in phosphate starvation-induced lipid remodeling, and are crucial for low-phosphate adaptation. *Plant J.* 2009;57:322–331
- Koenen EJM, Ojeda DI, Steeves R, Migliore J, Bakker FT, Wieringa, JJ, Kidner C, Hardy OJ, Pennington RT, Bruneau A, Hughes CE. 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 Phytol.* 2020;225(3):1355-1369.
- Krasensky J, Jonak C. Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J Exp Bot.* 2012;63(4):1593-1608. doi:10.1093/jxb/err460.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* 2001;305:567-580.
- Kuznetsova I, Lugmayr A, Siira SJ, Racham O, Filipovska A. CirGO: an alternative circular way of visualising gene ontology terms. *BMC Bioinformatics* 2019;20:84. doi:10.1186/s12859-019-2671-2
- Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 2007;35(9):3100–3108.
- Lanes EC, Pope NS, Alves R, Carvalho Filho NM, Giannini TC, Giulietti AM, et al. Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from Amazonian Savannas. *Front. Plant Sci.* 2018;9:532. doi:10.3389/fpls.2018.00532
- Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* 2012;9:357-359.
- Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.
- Li M, Cao L, Mwimba M, Zhou Y, Li L, Zhou M, Schnable PS, O'Rourke JA, Dong X, Wang W. Comprehensive mapping of abiotic stress inputs into the soybean circadian clock. *PNAS.* 2019;116(47):23840–23849; doi:10.1073/pnas.1708508116.
- Li M-W, Lam H-M. The Modification of Circadian Clock Components in Soybean During Domestication and Improvement. *Front. Genet.*, 2020; 11:571188; doi: 10.3389/fgene.2020.571188
- Liu Q, Luo L, Zheng L. Lignins: Biosynthesis and Biological Functions in Plants. *Int. J. Mol. Sci.* 2018;19:335.
- Long RW, Bush SE, Grady KC, Smith DS, Potts DL, D'Antonio CM, Dudley TL, Fehlberg SD, Gaskin JF, Glenn EP, Hultine KR. Can local adaptation explain varying patterns of herbivory tolerance in a recently introduced woody plant in North America? *Conservation Physiology.* 2017;5(1):cox016; doi:10.1093/conphys/cox016.

Lorenzi H. Árvores brasileiras: manual de identificação e cultivo de plantas arbóreas nativas do Brasil. São Paulo: Instituto Plantarum. 2002;2:179-187.

LPWG (Legume Phylogeny Working Group). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon*. 2017;66:44–77.

Lu BX, Zeng ZB, Shi TL. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci China Life Sci*. 2013;56:143–155; doi:10.1007/s11427-013-4442-z.

Mamrot J. *et al.* De novo transcriptome assembly for the spiny mouse (*Acomys cahirinus*). *Scientific Reports*. 2017;7 doi:10.1038/s41598-017-09334-7.

Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*. 2005;21(19):3787–3793.

Mattos CMJ, Silva WLS, Carvalho CS, Lima AN, Faria SM, Lima HC. Flora das cangas da serra dos Carajás, Pará, Brasil: Leguminosae. *Rodriguésia*. 2018;69(3):1147-1220.

McClung CR. The plant circadian oscillator. *Biology*. 2019;8(1):14; doi:10.3390/biology8010014.

Mcdonald, T.; Jonson, J.; Dixon, K.W. National standards for the practice of ecological restoration in Australia. *Restor. Ecol*. 2016;24:4–32.

Moreno-Santillán DD, Machain-Williams C, Hernández-Montes G, Ortega J. De Novo Transcriptome Assembly and Functional Annotation in Five Species of Bats. *Sci. Rep*. 2019;9:6222; doi:10.1038/s41598-019-42560-9

Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS ONE*. 2014;9(3):e91776.

Nielsen H. Predicting Secretory Proteins with SignalP. *Methods Mol Biol*. 2017;1611:59-73. doi: 10.1007/978-1-4939-7015-5\_6.

Nishad R, Ahmed T, Rahman VJ, Kareem A. Modulation of Plant Defense System in Response to Microbial Interactions. *Front Microbiol*. 2020; 11:1298; doi:10.3389/fmicb.2020.01298

Ntiamao-Baidu Y. Can wildlife contribute to food security in Africa? Issues and conclusions. In: Ntiamao-Baidu editor. *Wildlife and food security in Africa*. 1997;91-99; ISBN 92-5-104103-2 (FAO of the United Nations).

Oellrich A, Sanger Mouse Genetics Project, Smedley D. Linking tissues to phenotypes using gene expression profiles. *Database*. 2014;bau017. doi:10.1093/database/bau017.

- Oliveira LC. Filogenia de *Parkia* R.Br. (Leguminosae: Mimosoideae) baseada em sequências de DNA de cloroplasto. Instituto Nacional Pesquisas da Amazônia – INPA. 2015.
- Ormeño E, Fernandez C. Effect of Soil Nutrient on Production and Diversity of Volatile Terpenoids from Plants. *Curr Bioact Compd.* 2012; 8(1):71-79; doi:10.2174/157340712799828188
- Paunov M, Koleva L, Vassilev A, Vangronsveld J, Goltsev V. Effects of different metals on photosynthesis: cadmium and zinc affect chlorophyll fluorescence in durum wheat. *Int. J. Mol. Sci.* 2018;19(3):787.
- Perea-García A, Andrés-Colás N, Peñarrubia L. Copper homeostasis influences the circadian clock in *Arabidopsis*. *Plant signaling & behavior.* 2010;5(10):1237–1240; doi:10.4161/psb.5.doi:10.12920.
- Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 2009;26:1641–1650.
- Ramos SJ, Caldeira CF, Gastauer M, Costa DLP, Furtini Neto AE, Souza FBM, Souza-Filho PWM, Siqueira JO. Native leguminous plants for mineland revegetation in the eastern Amazon: seed characteristics and germination. *New Forests.* 2019;50:859–872; doi:10.1007/s11056-019-09704-1.
- Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 2013; 14(9):R95. doi:10.1186/gb-2013-14-9-r95
- Rivers MC, Brummitt NA, Nic Lughadha E, Meagher TR. Do species conservation assessments capture genetic diversity? *Glob. Ecol. Conserv.* 2014;2:81–87.
- Roberts WR, Roalson EH. Comparative transcriptome analyses of flower development in four species of *Achimenes* (Gesneriaceae). *BMC Genomics.* 2017;18(1):240; doi:10.1186/s12864-017-3623-8.
- Robinson MD, McCarthy DJ, Smyth GK. EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics.* 2010;26(1):139–40.
- Salgado, AAR, Carmo, FF. *World Geomorphological Landscapes.* Springer; New York: *Quadrilátero Ferrífero: A Beautiful and Neglected Landscape Between the Gold and Iron Ore Reservoirs.* 2015;319–330.
- Schaefer CE, Cândido HG, Corrêa GR, Pereira A, Nunes JA, Souza OF, Marins A, Filho EF, Ker JC. Solos desenvolvidos sobre canga ferruginosa no Brasil: uma revisão crítica e papel ecológico de termiteiros. In: Carmo FF, Kamino LHY, editors. *Geossistemas Ferruginosos do Brasil: áreas prioritárias para conservação da diversidade geológica e biológica, patrimônio cultural e serviços ambientais.* 3i; Belo Horizonte: 77-102., 2015
- Schulte PM. Changes in gene expression as biochemical adaptations to environmental change: a tribute to Peter Hochachka. *Comparative biochemistry and physiology. Part Molecular Biology* 2004;139:519–529.

Schulz, M. H., Zerbino D. R., Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012;28(8):1086–1092; doi:10.1093/bioinformatics/bts094.

Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*. 2016;22(6):839-851. doi:10.1261/rna.053959.115

Schlichting CD, Smith H. Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. *Evolutionary Ecology*. 2002;16:189–211.

Silva MFF. Distribuição de metais pesados na vegetação metalófila de Carajás. *Acta Bot. Bras.* 1992;6, 107–22.

Silva JS, Gastauer M, Ramos SJ, Mitre S, Neto AEF, Siqueira O, Caldeira CF. Initial growth of Fabaceae species: Combined effects of topsoil and fertilizer application for mineland revegetation. *Flora*. 2018;246 doi:10.1016/j.flora.2018.08.001.

Silveira F. A. O., Negreiros D., Barbosa N. P. U. et al. Ecology and evolution of plant diversity in the endangered campo rupestre: a neglected conservation priority. *Plant Soil* 2016;1:129–52.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.

Simon MF, Pastore JFB, Souza AF, Borges LM, Scalon V, Ribeiro PG, Silva JS, Souza V, Queiroz L. Molecular phylogeny of *Stryphnodendron* (Mimosoideae, Leguminosae) and generic delimitations in the Piptadenia group. *International Journal of Plant Sciences* 2016;177:44–59.

Simmons GC. Canga Caves in the Quadrilátero Ferrífero, Minas Gerais, Brazil. *The National Speleological Society Bulletin*, 1963;25(1):66-72.

Sinha R, Gupta A, Senthil-Kumar M. Concurrent Drought Stress and Vascular Pathogen Infection Induce Common and Distinct Transcriptomic Responses in Chickpea. *Front. Plant Sci.* 2017; doi: 10.3389/fpls.2017.00333.

Skirycz A, Castilho A, Chaparro C, Carvalho N, Tzotzos G, Siqueira JO. Canga biodiversity, a matter of mining. *Front. Plant Sci.* 2014;5:1-9.

Sonter LJ, Barrett DJ, Soares-Filho BS, Moran CJ. Global demand for steel drives extensive land-use change in Brazil's Iron Quadrangle. *Global Environmental Change*. 2014;26:63–72.

Sork VL, Aitken SN, Dyer RJ, Eckert AJ, Legendre P, Neale DB. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes* 2013;9:901-911.

Souza FCR, Carmo FF. Geossistemas Ferruginosos no Brasil. In: Flávio Fonseca do Carmo; Luciana Hiromi Yoshino Kamino. (Org.). Geossistemas Ferruginosos do Brasil: áreas prioritárias para conservação da diversidade geológica e biológica, patrimônio cultural e serviços ambientais. 1ed. Belo Horizonte: 3i Editora, 2015;1:47-76.

Souza-Filho PWM, Giannini TC, Jaffé R, Giulietti AM, et al. Mapping and quantification of ferruginous outcrop savannas in the Brazilian Amazon: A challenge for biodiversity conservation. *PloS One*. 2019;14(1), e0211095.

Spillane JL, LaPolice TM, MacManes, MD, Plachetzki DC, Signal, bias, and the role of transcriptome assembly quality in phylogenomic inference. Preprint at bioRxiv. 2020; doi:10.1101/2020.07.23.216606.

Steed G, Ramirez DC, Hannah MA, Webb AAR. Chronoculture, harnessing the circadian clock to improve crop yield and sustainability. *Science*, 2021;372, eabc9141. doi: 10.1126/science.abc9141.

Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLOS ONE* 2011;6(7): e21800; doi:10.1371/journal.pone.0021800

Tahmasebi A, Ashrafi-Dehkordi E, Shahriari AG, Mazloomi SM, Ebrahimie E. Integrative meta-analysis of transcriptomic responses to abiotic stress in cotton. *Progress in Biophysics and Molecular Biology* 2019;146:112-122.

Tedersoo L, Laanisto L, Rahimlou S, Toussaint A, Hallikma T, Pärtel M. Global database of plants with root-symbiotic nitrogen fixation: NodDB. *Journal of Vegetation Science*. 2018;29(3):560-568; doi:10.1111/jvs.12627.

Torales SL, Rivarola M, Pomponio MF, et al. De novo assembly and characterization of leaf transcriptome for the development of functional molecular markers of the extremophile multipurpose tree species *Prosopis alba*. *BMC Genomics*. 2013;14:705; doi:10.1186/1471-2164-14-705.

Tunc-Ozdemir M, Miller G, Song L, et al. Thiamin confers enhanced tolerance to oxidative stress in *Arabidopsis*. *Plant Physiol*. 2009;151(1):421-432. doi:10.1104/pp.109.140046.

Van Der Bij AU, Weijters MJ, Robbink R, Harris J, Pawlett M, Ritz K, Benetková P, Moradi J, Frouz J, Diggelen R. Facilitating ecosystem assembly: Plant-soil interactions as a restoration tool. *Biological Conservation*. 2018; 220. 10.1016/j.biocon.2018.02.010.

Vance CP, Uhde-Stone C, Allan DL. Phosphorus acquisition and use: Critical adaptations by plants for securing a nonrenewable resource. *New Phytol*. 2003;157:423–447.

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience*. 2018;7(8):1-15; doi:10.1093/gigascience/giy093.

Viana, P. L., Mota, N. F. O., Gil, A. S. B, et al. Flora of the cangas of the Serra dos Carajás, Pará, Brazil: history, study and methodology. *Rodriguésia* 2016; 67(5): 1107-1124.

- Viehweger, K. How plants cope with heavy metals. *Bot Stud* 2014; 55:35. doi:10.1186/1999-3110-55-35
- Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* 2015;31(17):2912-4. doi:10.1093/bioinformatics/btv300
- Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group, 2009;10(1):57–63.
- Wang GY, Shi JL, Ng G, Battle SL, Zhang C, Lu H. Circadian clock-regulated phosphate transporter PHT4;1 plays an important role in *Arabidopsis* defense. *Mol Plant*. 2011;4(3):516-526. doi:10.1093/mp/ssr016.
- Wang YH, Qu XJ, Chen SY, Li DZ, Yi TS. Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. *Tree Genetics & Genomes*. 2017;13:41.
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. 2016. <https://ggplot2.tidyverse.org>.
- Willoughby J. R., Sundaram M., Wijayawardena B. K., Kimble S. J. A., Ji Y., Fernandez N. B., et al. The reduction of genetic diversity in threatened vertebrates and new recommendations regarding IUCN conservation rankings. *Biol. Conserv.* 2015;191:495–503.
- Xie Y, Wu G, Tang J, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2013;30(12):1660–1666. doi:10.1093/bioinformatics/btu077.
- Yang S, Yi K, Chang MM, Ling GZ, Zhao ZK, Li XF. Sequestration of Mn into the cell wall contributes to Mn tolerance in sugarcane (*Saccharum officinarum* L.). *Plant Soil* 2019;436:475–487.
- Yerushalmi S, Green RM. Evidence for the adaptive significance of circadian rhythms. *Ecology Letters* 2009;12(9):970-981.
- Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):14.
- Yu B, Xu C, Benning C. *Arabidopsis* disrupted in SQD2 encoding sulfolipid synthase is impaired in phosphate-limited growth. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(8):5732-5737. DOI: 10.1073/pnas.082696499
- Zahran HH. Rhizobium-legume symbiosis and nitrogen fixation under severe conditions and in an arid climate. *Microbiol Mol Biol Rev*. 1999;63(4):968-989.
- Zamioudis C, Pieterse CMJ. Modulation of host immunity by beneficial microbes. *Mol. Plant Microb. Inter*. 2012; 25:139-150. 10.1094/mpmi-06-11-0179

Zhang J, Ruhlman TA, Mower JP, Jansen RK. Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. *BMC Plant Biol.* 2013;13:228. doi:10.1186/1471-2229-13-228

## 8. Considerações finais

Comecei a trabalhar no Instituto Tecnológico Vale como bolsista de Desenvolvimento Tecnológico Industrial logo que terminei o mestrado, ainda antes de decidir o que faria no doutorado. Inicialmente trabalhei gerando códigos de barra de DNA para as plantas das cangas de Carajás. Trabalho que posteriormente se estendeu para invertebrados de cavernas e envolvia extração de DNA e sequenciamento Sanger de centenas de amostras. Foi a primeira grande diferença que percebi entre trabalhar na indústria e na academia: o volume. Volume de amostras (por semana, na época, extração de DNA de pelo menos 96 amostras e produção de até 12 placas de sequenciamento), volume de metas, de projetos, volume de treinamentos e volume de oportunidades. Entrar no ITV me deu a oportunidade de participar de diversos projetos de pesquisa, entender o funcionamento e um pouco do pensamento de uma indústria, participar de treinamentos, aprender como funcionam vários equipamentos que usamos na pesquisa e no cotidiano de laboratório. Fui encorajada a trabalhar em diferentes funções, na bancada, na análise de dados, no gerenciamento de amostras, reagentes e equipamentos, o que aumentou meu conjunto de habilidades.

O tipo de descoberta que desenvolvemos com pesquisa básica tem amplo impacto no conhecimento científico e é de extrema importância para a sociedade. No entanto, para mim sempre foi um desafio pensar na futura aplicação da pesquisa que desenvolvia. Trabalhar na iniciativa privada me ajudou a pensar em aplicações mais claras para as pesquisas, mesmo que a longo prazo. O ambiente é extremamente colaborativo e me deu a oportunidade de trabalhar em estreita interação com especialistas de outras áreas do conhecimento, decidindo coletivamente sobre as escolhas integrativas ideais para um determinado projeto ou meta. Essas colaborações são extremamente gratificantes porque juntos somos capazes de fazer algo que nunca faríamos sozinhos. Além disso, por via da empresa que nos mantém, temos a chance de aproximar a conversa entre indústria e academia e a vivência da comunicação científica, desde para pequenos produtores e cooperativas até o próprio meio científico e os cargos mais altos das empresas.

Nas próximas sessões apresento as premiações e os artigos, publicados durante o período de desenvolvimento da tese, relacionados aos outros projetos de pesquisa que participei no ITV.

## 8.1 Premiação

**Menção Honrosa** – Prêmio Alcides Carvalho na área de Genética Vegetal do Congresso Brasileiro de Genética 2021. Trabalho intitulado "Cross-species transcriptomes comparison reveals molecular signatures for plants development on iron-rich rocky outcrops".

## 8.2 Premiação suplementar

**Menção Honrosa** – RNA and Transcriptomics no X-meeting eXperience 2020. Trabalho intitulado "De novo transcriptomes of endangered Amazonian bats and the detection of a rich source of virus information".

## 8.3 Produção científica suplementar no período de doutoramento

### 1. Unraveling the plant diversity of the Amazonian canga through DNA barcoding.

Santelmo Vasconcelos, Gisele L. Nunes, **Mariana C. Dias**, Jamily Lorena, Renato R. M. Oliveira, Talvâne G. L. Lima, Eder S. Pires, Rafael B. S. Valadares, Ronnie Alves, Maurício T. C. Watanabe, Daniela C. Zappi, Alice L. Hiura, Mayara Pastore, Liziane V. Vasconcelos, Nara F. O. Mota, Pedro L. Viana, André S. B. Gil, André O. Simões, Vera L. Imperatriz-Fonseca, Raymond M. Harley, Ana M. Giulietti, Guilherme Oliveira 

*Ecology and Evolution*. 2021; 11:13348–13362

### 2. Conservation implications of genetic structure in the narrowest endemic quillwort from the Eastern Amazon

Jeronymo Dalapicolla, Ronnie Alves, Rodolfo Jaffé, Santelmo Vasconcelos, Eder Soares Pires, Gisele Lopes Nunes, Jovani Bernardino de Souza Pereira, José Tasso F. Guimarães, **Mariana C. Dias**, Taís Nogueira Fernandes, Daniela Scherer, Fernando Marino Gomes dos Santos, Alexandre Castilho, Mirella Pupo Santos, Emiliano Nicolas Calderón, Rodrigo Lemes Martins, Rodrigo Nunes da Fonseca, Francisco de Assis Esteves, Cecílio Frois Caldeira, Guilherme Oliveira 

*Ecology and Evolution*. 2021; 11:10119–10132.

### 3. Complete mitochondrial genome of a troglophile Cydnidae (Hemiptera)

**Mariana Costa Dias**, Renato Renison Moreira Oliveira, Santelmo Vasconcelos, Eder Soares Pires, Xavier Prous, Thadeu Pietrobon, Guilherme Oliveira 

*Mitochondrial DNA Part B*. 2019, Vol. 4, No. 1, 420-422

**4. Quillworts from the Amazon: A multidisciplinary populational study on *Isoetes serracarajensis* and *Isoetes cangae***

Gisele Lopes Nunes, Renato Renison Moreira Oliveira, José Tasso Felix Guimarães, Ana Maria Giulietti, Cecílio Caldeira, Santelmo Vasconcelos, Eder Pires, **Mariana Dias**, Maurício Watanabe, Jovani Pereira, Rodolfo Jaffé, Cinthia Helena M. M. Bandeira, Nelson Carvalho-Filho, Edilson Freitas da Silva, Tarcísio Magevski Rodrigues, Fernando Marino Gomes dos Santos, Taís Fernandes, Alexandre Castilho, Pedro Walfir M. Souza-Filho, Vera Imperatriz-Fonseca, José Oswaldo Siqueira, Ronnie Alves, Guilherme Oliveira 

*PLoS ONE*. 2018; 13 (8): e0201417

**5. Blind Testing: DNA Barcoding Sheds Light Upon the Identity of Plant Fragments as a Subsidy for Cave Conservation**

Aline J. Ramalho, Daniela C. Zappi, Gisele L. Nunes, Mauricio T. C. Watanabe, Santelmo Vasconcelos, **Mariana C. Dias**, Rodolfo Jaffé, Xavier Prous, Tereza C. Giannini, Guilherme Oliveira, Ana M. Giulietti 

*Front. Plant Sci*. 2018; 9:1052

**6. Natural history of the narrow endemics *Ipomoea cavalcantei* and *I. marabaensis* from Amazon Canga savannahs**

Elena Babychuk, Sergei Kushnir, Santelmo Vasconcelos, **Mariana Costa Dias**, Nelson Carvalho-Filho, Gisele Lopes Nunes, Jorge Filipe dos Santos, Lourival Tyski, Delmo Fonseca da Silva, Alexandre Castilho, Vera Lucia Imperatriz Fonseca & Guilherme Oliveira 

*Scientific Reports*. 2017; 7:7493

## 9. APÊNDICES

### 9.1 APÊNDICE 1: Lista completa de Termos GO enriquecidos nos GDE de *P. platycephala* cultivada em substrato de canga.

GO	Nº de genes	Termo	Categoria	FDR	GDEs
GO:0048511	23	rhythmic process	BP	1.10E-06	Up
GO:0007623	15	circadian rhythm	BP	0.01349767	Up
GO:0000272	12	polysaccharide catabolic process	BP	0.01491729	Up
GO:0050896	157	response to stimulus	BP	0.01491729	Up
GO:0051707	53	response to other organism	BP	0.01491729	Up
GO:0009266	36	response to temperature stimulus	BP	0.01491729	Up
GO:0009628	81	response to abiotic stimulus	BP	0.01491729	Up
GO:0044247	7	cellular polysaccharide catabolic process	BP	0.01491729	Up
GO:0018131	3	oxazole or thiazole biosynthetic process	BP	0.01491729	Up
GO:0046484	3	oxazole or thiazole metabolic process	BP	0.01491729	Up
GO:0052837	3	thiazole biosynthetic process	BP	0.01491729	Up
GO:0052838	3	thiazole metabolic process	BP	0.01491729	Up
GO:0009409	27	response to cold	BP	0.01677381	Up
GO:0009607	54	response to biotic stimulus	BP	0.02154973	Up
GO:0043207	53	response to external biotic stimulus	BP	0.02154973	Up
GO:0005618	26	cell wall	CC	0.02154973	Up
GO:0030312	26	external encapsulating structure	CC	0.02154973	Up
GO:0009251	7	glucan catabolic process	BP	0.0229043	Up
GO:1901700	61	response to oxygen-containing compound	BP	0.02347917	Up
GO:0042752	10	regulation of circadian rhythm	BP	0.02453782	Up
GO:0009416	36	response to light stimulus	BP	0.0263975	Up
GO:2000028	10	regulation of photoperiodism, flowering	BP	0.02696027	Up
GO:0044275	8	cellular carbohydrate catabolic process	BP	0.03228892	Up
GO:0005983	5	starch catabolic process	BP	0.03747315	Up
GO:0042221	87	response to chemical	BP	0.04101128	Up
GO:0033993	31	response to lipid	BP	0.04382243	Up
GO:0009605	63	response to external stimulus	BP	0.04382243	Up
GO:0009314	36	response to radiation	BP	0.04382243	Up
GO:0044434	143	chloroplast part	CC	5.50E-13	Down
GO:0044435	143	plastid part	CC	5.50E-13	Down
GO:0009416	82	response to light stimulus	BP	7.97E-11	Down
GO:0009314	82	response to radiation	BP	5.00E-10	Down
GO:0009628	164	response to abiotic stimulus	BP	9.58E-10	Down
GO:0009507	154	chloroplast	CC	3.54E-09	Down
GO:0031969	44	chloroplast membrane	CC	2.61E-08	Down
GO:0042170	44	plastid membrane	CC	1.12E-07	Down
GO:0009526	59	plastid envelope	CC	1.32E-07	Down
GO:0009941	58	chloroplast envelope	CC	2.52E-07	Down
GO:0044436	62	thylakoid part	CC	1.13E-06	Down
GO:0034357	59	photosynthetic membrane	CC	1.13E-06	Down
GO:0042651	59	thylakoid membrane	CC	1.13E-06	Down

GO:0009536	167	plastid	CC	2.35E-06	Down
GO:0007623	25	circadian rhythm	BP	6.77E-06	Down
GO:0009535	56	chloroplast thylakoid membrane	CC	7.72E-06	Down
GO:0055035	56	plastid thylakoid membrane	CC	7.72E-06	Down
GO:0048511	28	rhythmic process	BP	1.33E-05	Down
GO:0031967	59	organelle envelope	CC	1.34E-05	Down
GO:0031975	59	envelope	CC	1.34E-05	Down
GO:0050896	284	response to stimulus	BP	2.01E-05	Down
GO:0009642	24	response to light intensity	BP	0.00011708	Down
GO:0009739	18	response to gibberellin	BP	0.00026598	Down
GO:0071483	9	cellular response to blue light	BP	0.00036093	Down
GO:0009637	17	response to blue light	BP	0.00117962	Down
GO:0009639	23	response to red or far red light	BP	0.00236787	Down
GO:0009644	14	response to high light intensity	BP	0.00236787	Down
GO:0042752	15	regulation of circadian rhythm	BP	0.00299864	Down
GO:0009534	27	chloroplast thylakoid	CC	0.00299864	Down
GO:0031976	27	plastid thylakoid	CC	0.00299864	Down
GO:0015075	52	ion transmembrane transporter activity	MF	0.00379884	Down
GO:0022857	70	transmembrane transporter activity	MF	0.00429484	Down
GO:0015979	21	photosynthesis	BP	0.00433963	Down
GO:0008509	26	anion transmembrane transporter activity	MF	0.0050833	Down
GO:0071482	17	cellular response to light stimulus	BP	0.00542259	Down
GO:0008878	5	glucose-1-phosphate adenylyltransferase activity	MF	0.00759375	Down
GO:0009651	47	response to salt stress	BP	0.00898856	Down
GO:0009532	53	plastid stroma	CC	0.00898856	Down
GO:0005215	73	transporter activity	MF	0.00913389	Down
GO:0009649	8	entrainment of circadian clock	BP	0.00956696	Down
GO:0010100	6	negative regulation of photomorphogenesis	BP	0.01292208	Down
GO:0055083	6	monovalent inorganic anion homeostasis	BP	0.01292702	Down
GO:0015740	6	C4-dicarboxylate transport	BP	0.01462963	Down
GO:0071478	17	cellular response to radiation	BP	0.01487636	Down
GO:0022891	57	NA	NA	0.01487636	Down
GO:0010207	8	photosystem II assembly	BP	0.01501103	Down
GO:0009624	15	response to nematode	BP	0.01501103	Down
GO:0009640	10	photomorphogenesis	BP	0.01715091	Down
GO:0009575	4	chromoplast stroma	CC	0.01715091	Down
GO:0042440	20	pigment metabolic process	BP	0.0175899	Down
GO:0005982	11	starch metabolic process	BP	0.0175899	Down
GO:0010114	13	response to red light	BP	0.02017416	Down
GO:0046364	10	monosaccharide biosynthetic process	BP	0.02069272	Down
GO:0009570	51	chloroplast stroma	CC	0.02069272	Down
GO:2000030	11	regulation of response to red or far red light	BP	0.02323889	Down
GO:0009706	15	chloroplast inner membrane	CC	0.0240235	Down
GO:0015291	28	secondary active transmembrane transporter activity	MF	0.02931906	Down
GO:0022892	60	NA	NA	0.02931906	Down
GO:0015743	5	malate transport	BP	0.02931906	Down

GO:0030388	5	fructose 1,6-bisphosphate metabolic process	BP	0.02931906	Down
GO:0006094	7	gluconeogenesis	BP	0.02931906	Down
GO:0006811	50	ion transport	BP	0.02931906	Down
GO:0044281	110	small molecule metabolic process	BP	0.02931906	Down
GO:0051241	20	negative regulation of multicellular organismal process	BP	0.02931906	Down
GO:0032922	5	circadian regulation of gene expression	BP	0.03086232	Down
GO:0015994	12	chlorophyll metabolic process	BP	0.0320746	Down
GO:0006820	27	anion transport	BP	0.0320746	Down
GO:0006835	7	dicarboxylic acid transport	BP	0.03326635	Down
GO:0010035	68	response to inorganic substance	BP	0.03358869	Down
GO:0055085	57	transmembrane transport	BP	0.03514264	Down
GO:0009528	15	plastid inner membrane	CC	0.0352261	Down
GO:0080167	17	response to karrikin	BP	0.0352261	Down
GO:0044711	104	NA	NA	0.03874438	Down
GO:0015711	21	organic anion transport	BP	0.03874438	Down
GO:0019252	8	starch biosynthetic process	BP	0.04022177	Down
GO:0010343	3	singlet oxygen-mediated programmed cell death	BP	0.04212304	Down
GO:0010099	8	regulation of photomorphogenesis	BP	0.04212304	Down
GO:0055062	5	phosphate ion homeostasis	BP	0.04212304	Down
GO:0072506	5	trivalent inorganic anion homeostasis	BP	0.04212304	Down
GO:0044710	202	NA	NA	0.04212304	Down
GO:0042754	5	negative regulation of circadian rhythm	BP	0.04355725	Down
GO:0019319	7	hexose biosynthetic process	BP	0.04432301	Down
GO:0006778	13	porphyrin-containing compound metabolic process	BP	0.04572035	Down
GO:0009579	31	thylakoid	CC	0.04592077	Down
GO:0008301	3	DNA binding, bending	MF	0.04631355	Down

**9.2 APÊNDICE 2:** Lista completa de Termos GO enriquecidos nos GDE de *S. pulcherrimum* cultivada em substrato de canga.

GO	Nº de GDEs	Termo	Categoria	FDR	GDEs
GO:0048511	15	rhythmic process	BP	0.01039796	Up
GO:0016115	9	terpenoid catabolic process	BP	0.00011096	Down
GO:0009628	123	response to abiotic stimulus	BP	0.00011096	Down
GO:0016036	15	cellular response to phosphate starvation	BP	0.00011096	Down
GO:0008300	9	isoprenoid catabolic process	BP	0.00013893	Down
GO:0031969	35	chloroplast membrane	CC	0.00013893	Down
GO:0044242	17	cellular lipid catabolic process	BP	0.00027033	Down
GO:0042910	12	xenobiotic transmembrane transporter activity	MF	0.00027033	Down
GO:0042170	35	plastid membrane	CC	0.00027033	Down
GO:0044255	55	cellular lipid metabolic process	BP	0.00030349	Down
GO:0009416	55	response to light stimulus	BP	0.00030349	Down
GO:0044710	180	NA	NA	0.00041563	Down
GO:0007623	20	circadian rhythm	BP	0.00059905	Down
GO:0009314	55	response to radiation	BP	0.00090307	Down
GO:0006629	64	lipid metabolic process	BP	0.00105556	Down
GO:0016107	5	sesquiterpenoid catabolic process	BP	0.00121579	Down
GO:0043290	5	apocarotenoid catabolic process	BP	0.00121579	Down
GO:0046345	5	abscisic acid catabolic process	BP	0.00121579	Down
GO:0031669	19	cellular response to nutrient levels	BP	0.00150191	Down
GO:0032104	6	regulation of response to extracellular stimulus	BP	0.00156218	Down
GO:0032107	6	regulation of response to nutrient levels	BP	0.00156218	Down
GO:0006721	21	terpenoid metabolic process	BP	0.00183177	Down
GO:0071483	8	cellular response to blue light	BP	0.00200782	Down
GO:0009639	20	response to red or far red light	BP	0.00224846	Down
GO:0071482	17	cellular response to light stimulus	BP	0.00224846	Down
GO:0031667	23	response to nutrient levels	BP	0.00233953	Down
GO:0044281	99	small molecule metabolic process	BP	0.0025611	Down
GO:0009637	15	response to blue light	BP	0.00275675	Down
GO:0071489	10	cellular response to red or far red light	BP	0.00275675	Down
GO:0042578	30	phosphoric ester hydrolase activity	MF	0.00275675	Down
GO:0016709	15	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	MF	0.00291958	Down
GO:0046164	8	alcohol catabolic process	BP	0.00307578	Down
GO:0044434	94	chloroplast part	CC	0.00307578	Down
GO:0022857	57	transmembrane transporter activity	MF	0.00307578	Down
GO:0006720	23	isoprenoid metabolic process	BP	0.00307578	Down
GO:0071478	17	cellular response to radiation	BP	0.00312086	Down
GO:0005215	61	transporter activity	MF	0.00312493	Down
GO:0044435	94	plastid part	CC	0.00318265	Down
GO:0016042	20	lipid catabolic process	BP	0.00318265	Down
GO:1901616	10	organic hydroxy compound catabolic process	BP	0.00322715	Down

GO:0016705	22	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	MF	0.00362688	Down
GO:0042594	20	response to starvation	BP	0.00499756	Down
GO:0071496	22	cellular response to external stimulus	BP	0.00632838	Down
GO:0016021	156	integral component of membrane	CC	0.00632838	Down
GO:0048511	20	rhythmic process	BP	0.00752716	Down
GO:0006066	15	alcohol metabolic process	BP	0.00752716	Down
GO:0031668	21	cellular response to extracellular stimulus	BP	0.00802797	Down
GO:1901615	26	organic hydroxy compound metabolic process	BP	0.00821718	Down
GO:0009991	25	response to extracellular stimulus	BP	0.00835065	Down
GO:0010017	9	red or far-red light signaling pathway	BP	0.00907848	Down
GO:0006820	23	anion transport	BP	0.00907848	Down
GO:0019374	5	galactolipid metabolic process	BP	0.00907848	Down
GO:0055062	6	phosphate ion homeostasis	BP	0.0091104	Down
GO:0072506	6	trivalent inorganic anion homeostasis	BP	0.0091104	Down
GO:0007154	23	cell communication	BP	0.00965048	Down
GO:0031224	158	intrinsic component of membrane	CC	0.00973313	Down
GO:0044711	91	NA	NA	0.01013904	Down
GO:0008194	20	UDP-glycosyltransferase activity	MF	0.01084474	Down
GO:0055081	10	anion homeostasis	BP	0.01127096	Down
GO:0009267	16	cellular response to starvation	BP	0.01127096	Down
GO:0003993	8	acid phosphatase activity	MF	0.01127096	Down
GO:0009687	6	abscisic acid metabolic process	BP	0.01127096	Down
GO:0043288	6	apocarotenoid metabolic process	BP	0.01127096	Down
GO:1902644	6	tertiary alcohol metabolic process	BP	0.01127096	Down
GO:0055085	48	transmembrane transport	BP	0.01127096	Down
GO:0016788	53	hydrolase activity, acting on ester bonds	MF	0.01189201	Down
GO:0004497	18	monooxygenase activity	MF	0.01274118	Down
GO:0016128	6	phytosteroid metabolic process	BP	0.01274118	Down
GO:0071456	18	cellular response to hypoxia	BP	0.01274118	Down
GO:0008509	19	anion transmembrane transporter activity	MF	0.0131059	Down
GO:0009820	9	alkaloid metabolic process	BP	0.01317279	Down
GO:0005975	48	carbohydrate metabolic process	BP	0.01317279	Down
GO:0036294	18	cellular response to decreased oxygen levels	BP	0.0133216	Down
GO:0004337	4	geranyltranstransferase activity	MF	0.01354354	Down
GO:0001666	20	response to hypoxia	BP	0.01363827	Down
GO:0071453	18	cellular response to oxygen levels	BP	0.01402955	Down
GO:0071214	22	cellular response to abiotic stimulus	BP	0.01485776	Down
GO:0032787	37	monocarboxylic acid metabolic process	BP	0.01485776	Down
GO:0044699	293	NA	NA	0.01546027	Down
GO:0010099	8	regulation of photomorphogenesis	BP	0.01644965	Down
GO:0005342	12	organic acid transmembrane transporter activity	MF	0.01644965	Down
GO:0046943	12	carboxylic acid transmembrane transporter activity	MF	0.01644965	Down
GO:0046505	3	sulfolipid metabolic process	BP	0.0165444	Down
GO:0046506	3	sulfolipid biosynthetic process	BP	0.0165444	Down
GO:0046715	3	active borate transmembrane transporter activity	MF	0.0165444	Down

GO:0008911	3	lactaldehyde dehydrogenase activity	MF	0.01692661	Down
GO:0010295	3	(+)-abscisic acid 8'-hydroxylase activity	MF	0.01718074	Down
GO:0036293	20	response to decreased oxygen levels	BP	0.01967283	Down
GO:0006855	7	drug transmembrane transport	BP	0.0212283	Down
GO:0016129	5	phytosteroid biosynthetic process	BP	0.02222331	Down
GO:0070482	20	response to oxygen levels	BP	0.02236954	Down
GO:0008559	5	xenobiotic transmembrane transporting ATPase activity	MF	0.02331184	Down
GO:0006714	6	sesquiterpenoid metabolic process	BP	0.02381529	Down
GO:0032106	4	positive regulation of response to extracellular stimulus	BP	0.02381529	Down
GO:0032109	4	positive regulation of response to nutrient levels	BP	0.02381529	Down
GO:0080040	4	positive regulation of cellular response to phosphate starvation	BP	0.02381529	Down
GO:0055083	6	monovalent inorganic anion homeostasis	BP	0.02395182	Down
GO:0044712	46	NA	NA	0.02477577	Down
GO:0015893	7	drug transport	BP	0.02478442	Down
GO:0006082	65	organic acid metabolic process	BP	0.02478442	Down
GO:0055114	65	oxidation-reduction process	BP	0.02478442	Down
GO:0016491	73	oxidoreductase activity	MF	0.02478442	Down
GO:0008610	41	lipid biosynthetic process	BP	0.02478442	Down
GO:0016791	24	phosphatase activity	MF	0.02512793	Down
GO:2000030	9	regulation of response to red or far red light	BP	0.02577387	Down
GO:0010100	5	negative regulation of photomorphogenesis	BP	0.02577979	Down
GO:0019375	4	galactolipid biosynthetic process	BP	0.02668021	Down
GO:0006006	9	glucose metabolic process	BP	0.02711736	Down
GO:0016114	16	terpenoid biosynthetic process	BP	0.02711736	Down
GO:0044723	35	NA	NA	0.02736812	Down
GO:0043436	64	oxoacid metabolic process	BP	0.02787175	Down
GO:0072505	6	divalent inorganic anion homeostasis	BP	0.02831753	Down
GO:0015238	5	drug transmembrane transporter activity	MF	0.02919376	Down
GO:0015849	13	organic acid transport	BP	0.02926605	Down
GO:0046942	13	carboxylic acid transport	BP	0.02926605	Down
GO:0015291	22	secondary active transmembrane transporter activity	MF	0.03287753	Down
GO:0080167	15	response to karrikin	BP	0.03469235	Down
GO:0098656	11	anion transmembrane transport	BP	0.03586649	Down
GO:0004365	4	glyceraldehyde-3-phosphate dehydrogenase (NAD <sup>+</sup> ) (phosphorylating) activity	MF	0.03613037	Down
GO:0043891	4	glyceraldehyde-3-phosphate dehydrogenase (NAD(P) <sup>+</sup> ) (phosphorylating) activity	MF	0.03613037	Down
GO:0008299	18	isoprenoid biosynthetic process	BP	0.03613037	Down
GO:0009739	11	response to gibberellin	BP	0.03732527	Down
GO:0006811	39	ion transport	BP	0.03764706	Down
GO:0046503	5	glycerolipid catabolic process	BP	0.03764706	Down
GO:0048878	26	chemical homeostasis	BP	0.03839677	Down
GO:0090484	5	NA	NA	0.03850132	Down
GO:0006835	5	dicarboxylic acid transport	BP	0.03850132	Down
GO:0006664	8	glycolipid metabolic process	BP	0.04129492	Down

GO:0046713	3	borate transport	BP	0.04249924	Down
GO:0008886	3	glyceraldehyde-3-phosphate dehydrogenase (NADP <sup>+</sup> ) (non-phosphorylating) activity	MF	0.04285656	Down
GO:0050896	207	response to stimulus	BP	0.04285656	Down
GO:0046486	16	glycerolipid metabolic process	BP	0.04285656	Down

**9.3 APÊNDICE 3:** Lista completa de Termos GO enriquecidos nos GDE de *P. platycephala* cultivada em estéril amarelo.

GO	Nº de GDEs	Termo	Categoria	FDR	GDEs
GO:0048511	17	rhythmic process	BP	7.96E-06	Down
GO:0042752	9	regulation of circadian rhythm	BP	0.00478126	Down
GO:0018131	3	oxazole or thiazole biosynthetic process	BP	0.00478126	Down
GO:0046484	3	oxazole or thiazole metabolic process	BP	0.00478126	Down
GO:0052837	3	thiazole biosynthetic process	BP	0.00478126	Down
GO:0052838	3	thiazole metabolic process	BP	0.00478126	Down
GO:0009228	4	thiamine biosynthetic process	BP	0.02576012	Down
GO:0042724	4	thiamine-containing compound biosynthetic process	BP	0.02576012	Down
GO:0006772	4	thiamine metabolic process	BP	0.0289709	Down
GO:0042723	4	thiamine-containing compound metabolic process	BP	0.0289709	Down
GO:0007623	10	circadian rhythm	BP	0.04172691	Down
GO:0009416	50	response to light stimulus	BP	2.96E-08	Up
GO:0009314	50	response to radiation	BP	5.83E-08	Up
GO:0007623	20	circadian rhythm	BP	7.65E-07	Up
GO:0031969	28	chloroplast membrane	CC	1.71E-06	Up
GO:0042170	28	plastid membrane	CC	3.71E-06	Up
GO:0009628	88	response to abiotic stimulus	BP	1.64E-05	Up
GO:0048511	20	rhythmic process	BP	1.81E-05	Up
GO:0050896	160	response to stimulus	BP	7.23E-05	Up
GO:0044434	62	chloroplast part	CC	0.00582059	Up
GO:0044435	62	plastid part	CC	0.00582059	Up
GO:0009706	12	chloroplast inner membrane	CC	0.00660771	Up
GO:0009941	29	chloroplast envelope	CC	0.00660771	Up
GO:0009526	29	plastid envelope	CC	0.00702372	Up
GO:0009528	12	plastid inner membrane	CC	0.00954183	Up
GO:0042752	10	regulation of circadian rhythm	BP	0.01745995	Up
GO:0008301	3	DNA binding, bending	MF	0.0306185	Up
GO:0071482	11	cellular response to light stimulus	BP	0.0306185	Up
GO:0010224	9	response to UV-B	BP	0.0306185	Up
GO:0009639	14	response to red or far red light	BP	0.03807403	Up
GO:0009411	12	response to UV	BP	0.03807403	Up
GO:0009637	10	response to blue light	BP	0.03807403	Up
GO:0004857	10	enzyme inhibitor activity	MF	0.04960552	Up

**9.4 APÊNDICE 4:** Lista completa de Termos GO enriquecidos nos GDE de *S. pulcherrimum* cultivada em estéril vermelho.

GO	Nº de GDEs	Termo	Categoria	FDR	GDEs
GO:0016036	12	cellular response to phosphate starvation	BP	0	Up
GO:0009267	13	cellular response to starvation	BP	6.04E-09	Up
GO:0031669	13	cellular response to nutrient levels	BP	6.34E-09	Up
GO:0042594	13	response to starvation	BP	9.74E-08	Up
GO:0007154	14	cell communication	BP	9.74E-08	Up
GO:0031668	13	cellular response to extracellular stimulus	BP	2.85E-07	Up
GO:0031667	13	response to nutrient levels	BP	3.22E-07	Up
GO:0071496	13	cellular response to external stimulus	BP	4.15E-07	Up
GO:0032104	5	regulation of response to extracellular stimulus	BP	1.08E-06	Up
GO:0032107	5	regulation of response to nutrient levels	BP	1.08E-06	Up
GO:0009991	13	response to extracellular stimulus	BP	4.66E-06	Up
GO:0003993	6	acid phosphatase activity	MF	5.39E-05	Up
GO:0032106	4	positive regulation of response to extracellular stimulus	BP	5.39E-05	Up
GO:0032109	4	positive regulation of response to nutrient levels	BP	5.39E-05	Up
GO:0080040	4	positive regulation of cellular response to phosphate starvation	BP	5.39E-05	Up
GO:0055062	5	phosphate ion homeostasis	BP	5.39E-05	Up
GO:0072506	5	trivalent inorganic anion homeostasis	BP	5.39E-05	Up
GO:0042578	12	phosphoric ester hydrolase activity	MF	6.53E-05	Up
GO:0055083	5	monovalent inorganic anion homeostasis	BP	8.00E-05	Up
GO:0072505	5	divalent inorganic anion homeostasis	BP	0.00016254	Up
GO:0055081	6	anion homeostasis	BP	0.00031769	Up
GO:0030643	3	cellular phosphate ion homeostasis	BP	0.00060964	Up
GO:0072501	3	cellular divalent inorganic anion homeostasis	BP	0.00060964	Up
GO:0072502	3	cellular trivalent inorganic anion homeostasis	BP	0.00060964	Up
GO:0016791	10	phosphatase activity	MF	0.00129831	Up
GO:0050801	9	ion homeostasis	BP	0.0016672	Up
GO:0030002	3	cellular anion homeostasis	BP	0.00208045	Up
GO:0030320	3	cellular monovalent inorganic anion homeostasis	BP	0.00208045	Up
GO:0008199	3	ferric iron binding	MF	0.00218843	Up
GO:0009247	4	glycolipid biosynthetic process	BP	0.0023708	Up
GO:0071456	8	cellular response to hypoxia	BP	0.00290947	Up
GO:0036294	8	cellular response to decreased oxygen levels	BP	0.00306621	Up
GO:0071453	8	cellular response to oxygen levels	BP	0.00316214	Up
GO:0098771	8	inorganic ion homeostasis	BP	0.00373051	Up
GO:0006664	4	glycolipid metabolic process	BP	0.00373051	Up
GO:1903509	4	liposaccharide metabolic process	BP	0.00612743	Up
GO:0001666	8	response to hypoxia	BP	0.00917461	Up

GO:0016788	14	hydrolase activity, acting on ester bonds	MF	0.00917461	Up
GO:0036293	8	response to decreased oxygen levels	BP	0.01144021	Up
GO:0070482	8	response to oxygen levels	BP	0.0117054	Up
GO:0033554	17	cellular response to stress	BP	0.01510524	Up
GO:0048878	9	chemical homeostasis	BP	0.01548047	Up
GO:0008198	3	ferrous iron binding	MF	0.02089734	Up
GO:0009605	20	response to external stimulus	BP	0.02110144	Up
GO:0046510	2	UDP-sulfoquinovose:DAG sulfoquinovosyltransferase activity	MF	0.02422212	Up
GO:0046475	2	glycerophospholipid catabolic process	BP	0.02889101	Up

**9.5 APÊNDICE 5:** Lista completa de Termos GO enriquecidos nos GDE de *S. pulcherrimum* cultivada em estéril amarelo.

GO	Nº de GDEs	Termo	Categoria	FDR	GDEs
GO:0009628	101	response to abiotic stimulus	BP	0.00051213	Up
GO:0044434	88	chloroplast part	CC	0.00874692	Up
GO:0009416	46	response to light stimulus	BP	0.00874692	Up
GO:0044435	88	plastid part	CC	0.00874692	Up
GO:0009314	46	response to radiation	BP	0.00934658	Up
GO:0071483	7	cellular response to blue light	BP	0.00934658	Up
GO:0042910	8	xenobiotic transmembrane transporter activity	MF	0.01009522	Up
GO:0032104	5	regulation of response to extracellular stimulus	BP	0.01009522	Up
GO:0032107	5	regulation of response to nutrient levels	BP	0.01009522	Up
GO:0031969	25	chloroplast membrane	CC	0.0191952	Up
GO:0071482	14	cellular response to light stimulus	BP	0.0191952	Up
GO:0050896	169	response to stimulus	BP	0.02175598	Up
GO:0071478	14	cellular response to radiation	BP	0.02175598	Up
GO:0008300	6	isoprenoid catabolic process	BP	0.02175598	Up
GO:0009820	9	alkaloid metabolic process	BP	0.02358561	Up
GO:0032106	4	positive regulation of response to extracellular stimulus	BP	0.02358561	Up
GO:0032109	4	positive regulation of response to nutrient levels	BP	0.02358561	Up
GO:0080040	4	positive regulation of cellular response to phosphate starvation	BP	0.02358561	Up
GO:0008509	15	anion transmembrane transporter activity	MF	0.02364254	Up
GO:0042170	25	plastid membrane	CC	0.02795777	Up
GO:0009639	16	response to red or far red light	BP	0.02950489	Up
GO:0015297	12	antiporter activity	MF	0.02950489	Up
GO:0009408	21	response to heat	BP	0.02950489	Up
GO:0006720	19	isoprenoid metabolic process	BP	0.02950489	Up
GO:0055081	9	anion homeostasis	BP	0.03001616	Up
GO:0071456	16	cellular response to hypoxia	BP	0.03432764	Up
GO:0036294	16	cellular response to decreased oxygen levels	BP	0.03738215	Up
GO:0001669	3	acrosomal vesicle	CC	0.03859379	Up
GO:0071453	16	cellular response to oxygen levels	BP	0.03886821	Up
GO:0009637	12	response to blue light	BP	0.0395059	Up
GO:0015291	18	secondary active transmembrane transporter activity	MF	0.04191038	Up
GO:0046686	25	response to cadmium ion	BP	0.04267147	Up
GO:0006721	16	terpenoid metabolic process	BP	0.04267147	Up
GO:0030643	3	cellular phosphate ion homeostasis	BP	0.04267147	Up
GO:0072501	3	cellular divalent inorganic anion homeostasis	BP	0.04267147	Up
GO:0072502	3	cellular trivalent inorganic anion homeostasis	BP	0.04267147	Up
GO:0016115	5	terpenoid catabolic process	BP	0.04267147	Up

GO:0055062	5	phosphate ion homeostasis	BP	0.04267147	Up
GO:0072506	5	trivalent inorganic anion homeostasis	BP	0.04267147	Up
GO:0009507	92	chloroplast	CC	0.04432666	Up
GO:0050474	3	(S)-norcochlorine synthase activity	MF	0.0465421	Up
GO:0009607	72	response to biotic stimulus	BP	2.07E-05	Down
GO:0043207	70	response to external biotic stimulus	BP	2.12E-05	Down
GO:0051707	68	response to other organism	BP	2.12E-05	Down
GO:0098542	53	defense response to other organism	BP	0.00017861	Down
GO:0009620	34	response to fungus	BP	0.00017861	Down
GO:0050832	29	defense response to fungus	BP	0.00017861	Down
GO:0051704	72	multi-organism process	BP	0.00017861	Down
GO:0006955	35	immune response	BP	0.00017861	Down
GO:0002376	40	immune system process	BP	0.00027331	Down
GO:0045087	33	innate immune response	BP	0.00079358	Down
GO:0006952	72	defense response	BP	0.00123716	Down
GO:0009605	79	response to external stimulus	BP	0.00211741	Down
GO:0005886	114	plasma membrane	CC	0.00551897	Down
GO:0031224	130	intrinsic component of membrane	CC	0.00551897	Down
GO:0005618	31	cell wall	CC	0.01550394	Down
GO:0030312	31	external encapsulating structure	CC	0.01550394	Down
GO:0019199	12	transmembrane receptor protein kinase activity	MF	0.02606915	Down
GO:0009617	37	response to bacterium	BP	0.02959176	Down
GO:0005576	47	extracellular region	CC	0.03013105	Down

## 9.6 APÊNDICE 6: Artigo publicado na revista *BMC Genomics* contendo as partes 1 e 2 do presente trabalho.

Dias et al. *BMC Genomics* (2022) 23:313  
<https://doi.org/10.1186/s12864-022-08449-0>

BMC Genomics

RESEARCH

Open Access



# Cross-species transcriptomes reveal species-specific and shared molecular adaptations for plants development on iron-rich rocky outcrops soils

Mariana Costa Dias<sup>1,2</sup>, Cecílio Caldeira<sup>1</sup>, Markus Gastauer<sup>1</sup>, Silvio Ramos<sup>1</sup> and Guilherme Oliveira<sup>1\*</sup> 

## Abstract

**Background:** *Canga* is the Brazilian term for the savanna-like vegetation harboring several endemic species on iron-rich rocky outcrops, usually considered for mining activities. *Parkia platycephala* Benth. and *Stryphnodendron pulcherrimum* (Willd.) Hochr. naturally occur in the *cangas* of Serra dos Carajás (eastern Amazonia, Brazil) and the surrounding forest, indicating high phenotypic plasticity. The morphological and physiological mechanisms of the plants' establishment in the *canga* environment are well studied, but the molecular adaptative responses are still unknown. To understand these adaptative responses, we aimed to identify molecular mechanisms that allow the establishment of these plants in the *canga* environment.

**Results:** Plants were grown in *canga* and forest substrates collected in the Carajás Mineral Province. RNA was extracted from pooled leaf tissue, and RNA-seq paired-end reads were assembled into representative transcriptomes for *P. platycephala* and *S. pulcherrimum* containing 31,728 and 31,311 primary transcripts, respectively. We identified both species-specific and core molecular responses in plants grown in the *canga* substrate using differential expression analyses. In the species-specific analysis, we identified 1,112 and 838 differentially expressed genes for *P. platycephala* and *S. pulcherrimum*, respectively. Enrichment analyses showed that unique biological processes and metabolic pathways were affected for each species. Comparative differential expression analysis was based on shared single-copy orthologs. The overall pattern of ortholog expression was species-specific. Even so, we identified almost 300 altered genes between plants in *canga* and forest substrates with conserved responses in the two species. The genes were functionally associated with the response to light stimulus and the circadian rhythm pathway.

**Conclusions:** Plants possess species-specific adaptative responses to cope with the substrates. Our results also suggest that plants adapted to both *canga* and forest environments can adjust the circadian rhythm in a substrate-dependent manner. The circadian clock gene modulation might be a central mechanism regulating the plants' development in the *canga* substrate in the studied legume species. The mechanism may be shared as a common mechanism to abiotic stress compensation in other native species.

**Keywords:** *Caesalpinioideae*, *Canga*, Comparative transcriptomics, De novo transcriptome, *Fabaceae*, Gene expression plasticity, Ironstone outcrops, *Leguminosae*, RNA-seq

\*Correspondence: [guilherme.oliveira@itv.org](mailto:guilherme.oliveira@itv.org)

<sup>1</sup> Instituto Tecnológico Vale, Rua Boaventura da Silva 955, Belém, Pará CEP 66055-090, Brazil

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.