

## Caracterização Funcional do Transcriptoma de Amendoim Forrageiro

Jônatas Chagas de Oliveira<sup>1</sup>, Eduardo Fernandes Formighieri<sup>2</sup>, Ana Letycia Basso Garcia<sup>3</sup>, Gabriel Rodrigues Alves Margarido<sup>4</sup> e Tatiana de Campos<sup>5</sup>

<sup>1</sup>Biólogo, doutor em Biodiversidade e Biotecnologia, técnico de laboratório da Universidade Federal do Acre, Rio Branco, AC.

<sup>2</sup>Engenheiro-agrônomo, doutor em Biologia Funcional e Molecular, pesquisador da Embrapa Agroenergia, Brasília, DF.

<sup>3</sup>Engenheira-agrônoma, doutoranda em Genética e Melhoramento de Plantas, Escola Superior de Agricultura Luiz de Queiróz, Piracicaba, SP.

<sup>4</sup>Engenheiro-agrônomo, doutor em Genética e Melhoramento de Plantas, professor da Escola Superior de Agricultura Luiz de Queiróz, Piracicaba, SP.

<sup>5</sup>Bióloga, doutora em Genética e Biologia Molecular, pesquisadora da Embrapa Acre, Rio Branco, AC.

**Resumo** – O uso do amendoim forrageiro em consórcios com gramíneas nas pastagens e como cobertura verde, consorciado com culturas comerciais, tem crescido nos últimos anos. A análise do genoma funcional permite a identificação de genes de interesse agrônômico. Assim, o objetivo deste trabalho foi realizar a anotação funcional de genes do transcriptoma de folhas de *Arachis pintoi*. Dos 98.432 transcritos analisados, 69% apresentaram correspondências com o banco de dados de proteínas do National Center of Biotechnology Information. As classes função molecular (36%) e processo biológico (35,8%) representaram a maioria dos termos de Gene Ontology atribuídos, enquanto o componente celular (28,2%) apresentou menor número. A análise de expressão diferencial identificou 1.550 e 1.357 genes com maior nível de expressão nas cultivares Amarillo e Belomonte, respectivamente. A análise de enriquecimento dos genes mostrou que 55,63% pertencem à classe componente celular, seguida por função molecular (26,48%) e processo biológico (20,89%). Esses resultados são o primeiro relato de anotação funcional de *A. pintoi* que irá fornecer uma importante fonte de informação para avanços nos estudos de expressão, silenciamento e edição gênica nos programas de melhoramento de *Arachis*.

Termos para indexação: anotação funcional, *Arachis pintoi*, RNA-seq.

### Introdução

A utilização de leguminosas como cobertura verde consorciada com culturas comerciais é uma importante estratégia à conservação e melhoria da qualidade do solo, pois contribui para a manutenção da sua umidade além da fixação biológica do nitrogênio. Dentre as leguminosas, o amendoim forrageiro (*Arachis pintoi* Krapov. & WC Greg.) tem recebido destaque por seu uso em pastagens consorciadas com gramíneas, auxiliando no expressivo aumento do ganho de peso em bovinos e redução do tempo de abate (Maia, 2018). Apesar disso, a quantidade de cultivares disponíveis e o custo das sementes têm limitado a sua implantação em larga escala.

A tecnologia de sequenciamento de RNA (RNA-seq) tem demonstrado ser uma boa alternativa, pois permite a busca por marcadores moleculares com menor custo, além da identificação de genes de interesse agrônômico mais facilmente, utilizando análises de expressão gênica (Wit et al., 2015). A anotação do genoma tem por objetivo fazer o levantamento e rotulação das suas características relevantes, especialmente aquelas relacionadas aos aminoácidos e proteínas com funções específicas, responsáveis por controlar uma determinada característica (Balbinot, 2020). Nesse cenário, foi criado o Gene Ontology (GO) Consortium, o qual originou um vocabulário

controlado aplicável a todos os eucariotos, visando superar a falta de interoperabilidade de bancos de dados genômicos causada pela divergência na nomenclatura de genes e proteínas. Cada gene ou proteína pode ser descrito por um número limitado de termos de vocabulário que se enquadra em uma das três categorias ou domínios GO (processo biológico, função molecular ou componente celular), de acordo com a função que o produto do gene identificado pode desempenhar (Ashburner et al., 2000). A categoria processo biológico refere-se a um objetivo biológico para o qual o gene ou produto do gene contribui. A função molecular é definida como a atividade bioquímica do produto de um gene, descrevendo apenas o que é feito sem especificar onde ou quando o evento realmente ocorre. A categoria componente celular refere-se ao lugar na célula onde o produto de um gene está ativo (Ashburner et al., 2000).

Os estudos com transcriptoma no gênero *Arachis* estão concentrados, principalmente, no amendoim cultivado (*Arachis hypogaea* L.) e as espécies mais próximas filogeneticamente. Diante disso, o objetivo deste trabalho foi realizar a anotação funcional de genes do transcriptoma foliar de *A. pintoi*.

## Material e métodos

As análises foram realizadas com o transcriptoma de folhas de *A. pintoi* desenvolvido por Oliveira (2020), o qual utilizou duas cultivares: Amarillo e Belomonte. A anotação funcional foi feita por meio da comparação do transcriptoma com o banco de dados não redundante de proteínas (nr) do National Center of Biotechnology Information (NCBI), utilizando a opção BlastX do pacote Blast+<sup>1</sup>, com as configurações padrão. Os resultados foram importados para o programa Blast2GO<sup>2</sup> para mapeamento e recuperação de GO e anotações de código de enzima exclusivo (EC) dos unigenes, aplicando os parâmetros padrão. Os termos GO recuperados foram classificados em três categorias: componentes celulares, funções moleculares e processos biológicos.

Para as análises de expressão diferencial e enriquecimento foram estimadas as contagens de genes com o programa Salmon (Patro et al., 2017) e avaliada a expressão em nível de gene, agrupando transcrições em genes usando o pacote R tximport (Soneson et al., 2015). Em seguida, a matriz de contagens foi filtrada para manter apenas genes observados em três ou mais réplicas biológicas e com contagens por milhão (CPM  $\geq$  2). Posteriormente, a normalização dos dados foi realizada com a abordagem Trimmed Mean of M-values (TMM). A matriz resultante foi utilizada para realizar a análise de expressão diferencial (ED) contrastando as cultivares Amarillo e Belomonte. As análises subsequentes foram realizadas aplicando o método Quasi-Likelihood, implementado no pacote edgeR (Robinson et al., 2010). Os genes diferencialmente expressos foram usados para realizar a análise de enriquecimento de termos GO com o pacote GOseq (Young et al., 2010).

## Resultados e discussão

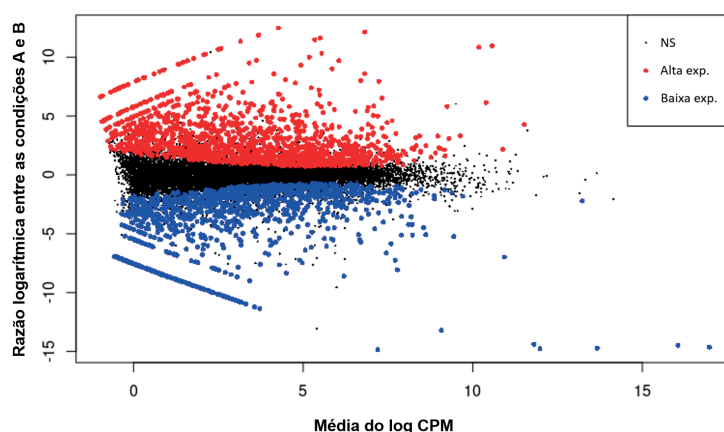
Entre os 98.432 transcritos analisados, 67.939 (69%) apresentaram homologia com sequências no banco de dados nr do NCBI. Esse banco de dados reúne um acervo de informações de bases tais como GenBank/GenPept, SwissProt, RefSeq, PIR, entre outras, as quais possuem grande quantidade de dados curados (alta qualidade) de sequências de proteínas. A anotação resultou em 37.219 transcritos com homologia com 146.844 termos GO, relativos às proteínas do banco de dados analisado. As classes função molecular (46.082 termos, 36%) e processo biológico

<sup>1</sup> Disponível em: <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>.

<sup>2</sup> Disponível em: <https://www.blast2go.com/>.

(45.800 termos, 35,8%) apresentaram a maioria dos termos GO atribuídos, enquanto em componente celular ocorreu o menor número (36.131 termos, 28,2%). O maior número de termos GO em relação ao número de transcritos deve-se à variação no comprimento dos transcritos analisados (180 a 16.665 pares de bases), o que permitiu identificar similaridade com mais de uma proteína em um mesmo transcrito. Além disso, como os termos GO são classificados de forma hierárquica, alguns deles podem estar relacionados com mais de uma categoria, uma vez que termos em níveis hierárquicos menores podem pertencer a mais de uma categoria hierárquica maior (Carnielle et al., 2015). Por outro lado, a ausência de termos GO em parte dos transcritos analisados não significa ausência de função, apenas que não há evidências do seu papel até o momento (Balbinot, 2020).

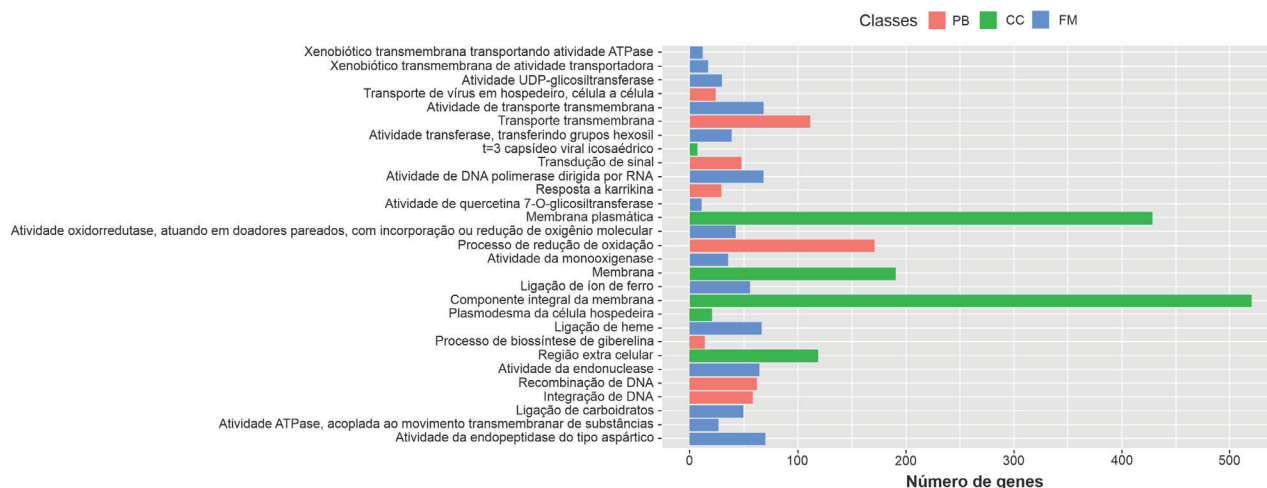
As cultivares Amarillo e Belomonte apresentaram 85,74% de expressão gênica semelhante no transcriptoma foliar. A análise de expressão diferencial encontrou 1.550 genes com maior nível de expressão em 'Amarillo' e 1.357 genes com maior nível de expressão em 'Belomonte' (Figura 1). Considerando os aspectos agrônômicos, as principais diferenças entre as cultivares são a quantidade de flores e sementes produzidas, velocidade de expansão e cobertura do solo e grau de tolerância à seca. Portanto, futuras análises dos genes diferencialmente expressos podem auxiliar na identificação daqueles responsáveis por regular características de interesse agrônômico e avaliar sua atuação no sistema foliar de *A. pintoi*.



**Figura 1.** Gráfico da análise da expressão diferencial das cultivares Amarillo e Belomonte de *Arachis pintoi*.

NS = Genes com expressão diferencial não significativa. Alta exp. = Genes com maior nível de expressão em 'Amarillo'. Baixa exp. = Genes com maior nível de expressão em 'Belomonte'. CPM = Contagem por milhão.

Como os termos GO podem apresentar redundâncias, a análise de enriquecimento é interessante por permitir a classificação funcional e a identificação dos genes mais representados nas amostras analisadas. Assim, a categoria componente celular (1.282 genes, 55,63%) foi a mais abundante, seguida por função molecular (645 genes, 26,48%) e processo biológico (509 genes, 20,89%) (Figura 2). Os termos GO componente integral da membrana (520 genes) e membrana plasmática (428 genes) foram os maiores grupos na classe componente celular. Na classe função molecular, atividade de endopeptidase do tipo aspártico (69 genes) e atividade de DNA polimerase dirigida por RNA (68 genes) foram as maiores, e processo de redução de oxidação (170 genes) e transporte transmembrana (111 genes) foram o primeiro e o segundo maior grupo em processos biológicos. Esses resultados apresentam um retrato dos processos e vias biológicas que provavelmente estão relacionados às condições biológicas das amostras no momento da coleta. A partir desses dados, o pesquisador poderá aplicar critérios de exclusão de termos que possuem uma relação fraca com os objetivos do estudo, baseado na pontuação do p-valor (Carnielle et al., 2015).



**Figura 2.** Distribuição de termos Gene Ontology enriquecidos no transcriptoma foliar de *Arachis pintoi*.

PB = Processo biológico. CC = Componente celular. FM = Função molecular.

Futuros estudos são necessários para elucidar o papel dos genes identificados em vias metabólicas relacionadas com características foliares do amendoim forrageiro, tais como aqueles responsáveis pelos mecanismos de ação na resistência a estresse hídrico. Essas informações são cruciais para o avanço do melhoramento de *Arachis*.

## Conclusões

As análises de anotação funcional do transcriptoma foliar de *A. pintoi* permitiram a identificação e classificação de genes.

## Agradecimento

Os autores agradecem o governo federal e governo do estado do Acre (Fapac TO: 024/2018) pelo apoio financeiro, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão de bolsa de produtividade e a Embrapa Acre pelo financiamento e infraestrutura para condução dos experimentos.

## Referências

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25-29, May 2000. DOI: <https://doi.org/10.1038/75556>.

BALBINOT, L. **Workflow científico de anotação genômica funcional e curadoria manual de genomas**. 2020. 82 f. Dissertação (Mestrado em Biotecnologia) – Universidade de Caxias do Sul, Caxias do Sul.

CARNIELLI, C. M.; WINCK, F. V.; LEME, A. F. P. Functional annotation and biological interpretation of proteomics data. **Biochimica et Biophysica Acta (BBA): Proteins and Proteomics**, v. 1854, n. 1, p. 46-54, Jan. 2015. DOI: <https://doi.org/10.1016/j.bbapap.2014.10.019>.

MAIA, G. F. N. **Desempenho produtivo de dois grupos genéticos de bovinos de corte em pastos puros e consorciados na Amazônia Ocidental**. 2018. 45 f. Dissertação (Mestrado em Ciência Animal) – Universidade Federal do Acre, Rio Branco.

OLIVEIRA, J. C. **Análise do genoma funcional de *Arachis pintoi* e desenvolvimento de novos marcadores moleculares**. 2020. 94 f. Tese (Doutorado em Biodiversidade e Conservação) – Universidade Federal do Acre, Rio Branco.

PATRO, R.; DUGGAL, G.; LOVE, M. I.; IRIZARRY, R. A.; KINGSFORD, C. Salmon provides fast and bias-aware quantification of transcript expression. **Nature Methods**, v. 14, n. 4, p. 417-419, Mar. 2017. DOI: <https://doi.org/10.1038/nmeth.4197>.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139-140, Jan. 2010. DOI: <https://doi.org/10.1093/bioinformatics/btp616>.

SONESON, C.; LOVE, M. I.; ROBINSON, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. **F1000Research**, v. 4, 1521, Dec. 2015. DOI: <https://doi.org/10.12688/f1000research.7563.2>.

WIT, P.; PESPENI, M. H.; PALUMBI, S. R. SNP genotyping and population genomics from expressed sequences – current advances and future possibilities. **Molecular Ecology**, v. 24, n. 10, p. 2310-2323, May. 2015. DOI: <https://doi.org/10.1111/mec.13165>.

YOUNG, M. D.; WAKEFIELD, M. J.; SMYTH, G. K.; OSHLACK, A. Gene ontology analysis for RNA-seq: accounting for selection bias. **Genome Biology**, v. 11, n. 2, R14, Feb. 2010. DOI: <https://doi.org/10.1186/gb-2010-11-2-r14>.