Weighted Gene Co-expression Network Analysis to Identify Key Modules and Hub Genes Related to Withanoloides Biosynthesis Pathway in *Datura metel*

Madhavi Hewadikaram[1,C], Kithmee De Silva[2]

[1,C] *NSBM Green University , Faculty of Science , Pitipana, Homagama, Sri Lanka*
madhavi.h@nsbm.ac.lk

[2]*Department of Plant Sciences, Faculty of Science, University of Colombo, Colombo 3, Sri Lanka*
desilvakithmee@gmail.com

ABSTRACT

The gene co-expression analysis and clustering concept were utilized to identify possible genes present in the withanolides biosynthesis pathway. Weighted Gene Co-expression Network Analysis (WCGNA) was carried out to identify putative genes. The identified known genes were used as bait to locate the co-expressed genes among them in identified modules in WGCNA analysis. As cytochrome P450s and glycosyltransferase (UGT) are candidates for withanolide biosynthesis in the literature, the study was focused on identifying CYPs and UGT genes. Finally, the study was able to locate 6 CYP genes and 1 UGT gene as putative candidate genes involved in the withanolides biosynthesis pathway. However, *in silico* characterization should be carried out for further isolation processes.

1. INTRODUCTION

Secondary metabolism facilitates primary metabolism and ensures that all metabolic pathways are assisted with the required regulation and coordination for the long-term persistence of the plant (Crozier *et al*., 2007; Nikolić *et al*., 2012). The resulting metabolites are secondary or specialized and are crucial components actively participating in plant defense mechanisms, hormonal regulation, enzyme catalytic functions, signaling coordination, and are even responsible for the plant's structure, vigour, and colour characteristics. Therefore, secondary metabolites are intermediate and promote growth and development but are not required for survival (Piasecka *et al*., 2015; Seca & Pinto, 2019).

Plant synthetic biology, transcriptomics, genomics, proteomics, metabolomics, and gene function prediction have all had a significant impact on identifying secondary metabolite biosynthetic pathways in plants (Zhang *et al*., 2017; Jacobowitz & Weng, 2020; Mutwil, 2020; Nützmann *et al*., 2016). These methods enabled the elucidation of unknown genes and enzymes in many plants' secondary metabolite biosynthetic pathways and *de novo* biosynthetic pathway prediction in many non-model plants (Tissier, 2012). The production of secondary metabolites is restricted to specialized organs, tissues, and different cell types. Moreover, these are regulated via different environmental parameters such as drought, salt resistance, or regulations via chemical elicitations like salicylic acid and methyl jasmonate (Wang *et al*., 2016). Therefore, the enzymes and their corresponding mRNAs should only be present in the cells where the specific metabolites are produced innately. These sets of genes co-regulate specific biological functions. This assumption has been exploited to identify biosynthetic genes by employing different gene expressions, gene co-expressions, and metabolite levels in omics technologies, and this is known as gene co-expression, which is based on the guilt by association principle (Gillis & Pavlidis, 2012; Saito *et al*., 2008; Yonekura-Sakakibara *et al*., 2008). This assumption was successfully proved

for fatty acid identification in tomato plants (Jeon *et al*., 2020). Once the known genes and their gene expressions are identified, they can be employed to uncover other unknown genes with a similar expression profile (Serin *et al*., 2016; Usadel *et al*., 2009). Several studies have pointed out that genes with similar expression patterns in different organs and developmental stages as well as biotic and abiotic agitations tend to involve similar biological processes (Dugé de Bernonville *et al*., 2017).

Gene clustering is highly employed in modern RNAseq studies (de Bernonville *et al*. 2020). Genes that form a cluster of genes display relative gene expression across the samples or conditions. In most gene clusters, genes responsible for specialized metabolites like secondary metabolites are located together in the genome. These were found in plants like Zea mays (Frey *et al*., 1997), rice (Yang *et al*., 2004), tomato (Akthar *et al*., 2013), and potato (Itkin *et al*., 2013). Therefore, the Gene Co-Expression Network Analysis (GCNA) method is commonly used in transcriptome-based gene co-expression networks (Higashi & Saito, 2013). It is a genetic approach to analyzing correlated and applicable genes in functional modules and phenotypic traits (Serin *et al*., 2016). There are numerous tools to analyze gene co-expression analysis. Of those, WGCNA is one of the most common GCNA-based approaches to identifying clustering genes in different metabolic pathways (Langfelder & Horvath, 2008). It is an R software package that uses a correlation of gene expressions for describing and visualizing data point networks related to gene networks (Langfelder & Horvath, 2008; Yao *et al.,* 2019). It is used to find biosynthesis pathways and genes that significantly impact those pathways. Moreover, this method can be employed to predict unknown genes and regulatory factors through the known genes in the networks. The crucial genes are named hub genes. *Modules* created in this WGCNA method are defined as clusters of highly interconnected genes. The number of genes found in whole genomic expressions is reduced by gene modules (Tai *et al*., 2018a). Each gene module consists of genes with the same function or similar biological regulation (Langfelder & Horvath, 2008). When the genes are annotated, like Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, that information can be related to modules to identify the hub genes and their modules. That would quickly lead to identifying interconnected and correlated modules. Hub genes enable the exploration of complex gene traits. This process will reduce the amount of data that needs to be processed, thus saving time. This method was extensively used in studies (DiLeo *et al*., 2011; Jia *et al*., 2021; B. Yu *et al*., 2020) and mainly in plants like tomato (DiLeo *et al*., 2011), soya bean (Gao *et al*., 2018), strawberry (Hollender *et al*., 2014) and tea (Tai *et al*., 2018). Moreover, Zheng *et al*. (2019) identified genes related to the expression pattern of the epidermal wax pathway in maize plants. This method was also utilized to identify hub genes related to abiotic drought stress-induced pathways in papaya plants (Gamboa-Tuz *et al*., 2018). Finally, this is one of the easiest methods to identify critical genes in interesting biological functions or synthesis pathways.

Withanolides are naturally occurring secondary metabolites that predominantly occur within plants of the *Solanaceae* family, especially the subfamily *Solanoidea* accounting for higher medicinal and economic importance (Chen *et al*., 2011; Dhar *et al*., 2015; Glotter, 1991). Withanolides are subdivided into nine withanolides, withaphysalins, physalins, nicandrenones, jaborols, ixocarpalactones, perulactones, acnistins and miscellaneous withasteroids (Misico *et al*., 2011). Withanolides are synthesized via mevalonate (MVA) and non-mevalonate (MEP/DOXP) pathways in the cytosol and plastids, respectively (Chaurasiya *et al*., 2012b; Dhar *et al*., 2015). *The Withanolide* biosynthesis pathway was elucidated up to 24-methyldesmosterol via 24-isomerase (EC 5.3.3), which catalyzes the conversion of 24-methylenecholesterol to 24-methyldesmosterol (Chaurasiya *et al*., 2012b; Gupta *et al*., 2013, 2015). No genes were identified after this step. Studies have pointed out that cytochrome P450 mono-oxygenase (CYP450s) and UDP-glycosyltransferases (UGTs) govern the

hydroxylation, oxidation, and glycosylation steps, yielding withanolides (Agarwal *et al*., 2017; Gupta *et al*., 2015; Senthil *et al*., 2015; Tripathi *et al*., 2016). However, the essential genes related to withanolide biosynthesis in *D. metel* have not yet been identified. Genes related to CYP450s and UGTs in plants are essential to the diversification of withanolides structures. RNA seq and other sequencing techniques provide a better opportunity to identify these possible candidates. Once these genes are identified, their synthesized proteins can be characterized via *in silico* and *in vivo* characterization methods. The objective of this paper is to isolate potential CYP450 and UGTs genes that might be involved in the biosynthesis pathway of withanolides in *D. metel* by using WGCNA analysis methods for the sequenced data.

2. METHODOLOGY

2.1 Transcript Quantification

Transcript-level quantification was performed to estimate gene and isoform expression levels from *D. metel* RNA-Seq data. The RSEM software package was used for this process. Initial sequencing reads and a *de novo* assembled transcriptome of *D. metel* were processed in this RSEM program to quantify the expression from transcriptome data. It aligned the reads against the reference annotated transcriptome and calculated relative abundances using the Bowtie2 aligner program within the RSEM program. The results provided the isoform level and gene level quantified tables called count tables

2.2 WGCNA Analysis

The WGCNA enables the identification of modules of highly correlated genes and hub genes with important effects in the withanolide biosynthesis pathway. Genes that were differentially expressed from the Differently Expressed Gene (DEG) analysis (7888 genes) were normalized, and their gene expression values (FPKM) were used as the input for WGCNA analysis in R software (version 4.0.5). WGCNA network construction and module detection were conducted using a signed topological overlap matrix (TOM). The soft power value is 18, the minimum module size is 30, and the merge cut height is 0.25. Each module was identified by color. In each module, the most significantly correlated genes with a WGCNA edge weight of 0.1 were visualized using Cytoscape 3.5.1 software. The most highly connected nodes within the module, known as "hub genes", were identified for each module. Functional analysis of the modules was carried out using Gene Ontology Enrichment. Further, the distribution of identified critical genes involved in the withanolides biosynthesis pathway in the modules was screened (Figure 01). Then the CYPs and UGTS co-expressed with known genes were identified as potential genes. KEGG pathway analysis was also performed on the modules.

2.3 Candidate gene selection based on the co-expressed gene expression criteria.

Due to the known localization of the metabolic pathway to leaf tissue, any gene not expressed in the leaf can be discarded. Moreover, the leaves elicited with SA allow us to identify the upregulated genes upon SA induction. The Withanolide biosynthetic pathway responds to SA elicitation (Dasgupta *et al.,* 2014; Sivanandhan *et al*., 2014), and therefore, genes that are preferentially expressed in response to SA elicitation are preferred candidates. The upregulated genes in the DEG analysis were further selected by co-expression analysis. Co-expression selection was based on gene discovery, utilizing the spatial proximity of genes on the genome that act as physical clustering of genes in secondary metabolism in plants. These methods are currently employed to identify candidate genes in specialized metabolism in several plant species (Nützmann *et al.,* 2016). Unigenes identified in each co-expression method were clustered together using a Venn diagram using Omics Box (2021). Unigenes co-expressed together or at least present in two co-expression methods were selected as the final possible candidate genes in the withanolide biosynthesis pathway for the *D.metel* transcriptome.

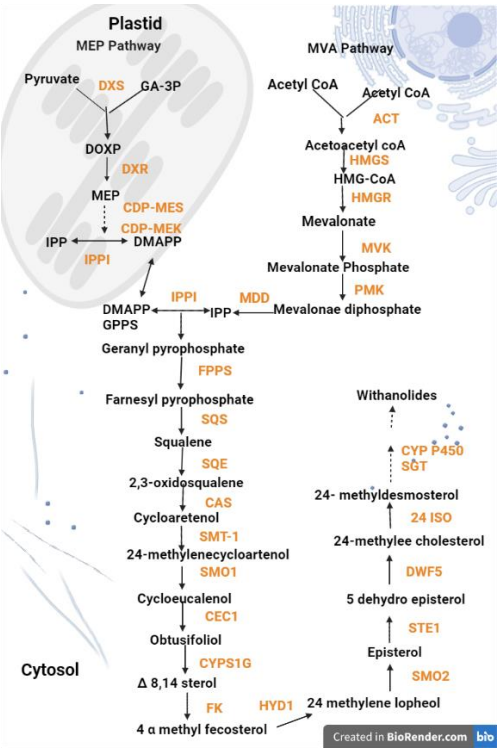

Figure 01: Known genes in Withanoloids Biosynthesis pathway

3.0 RESULTS AND DISCUSSION

There were 6 gene co-expression modules (Figure 02) identified under WGCNA method. The number of genes in the module ranged from 125 (Red) to -3403 (Turquoise). The cluster dendrogram generated (Figure 03) represents how all DE genes are distributed among modules. Each branch of the clustering tree represents a module with a different colour. Each leaf displays a gene. The correlation coefficient of the modules with each other represented the module relationships (Figure 04). Known genes present in the withanolide pathway were assessed, and they were mainly distributed to yellow, blue, and brown modules, respectively. Once the known genes were identified, CYP450s and UGTs were identified that co-expressed with the known genes in each module (Table 01).

Moreover, enriched CYP450 and UDP-GT were also investigated in three interesting modules. Unigenes that encode for CYP450, and UDP-GTs and their modules are given in Table 01. Their gene expressions were evaluated, and higher gene expressions were selected as candidate genes for withanolide production. KEGG pathway analysis revealed that brown and yellow modules consist of steroid and brassinosteroid, and triterpenoid biosynthesis pathways. Therefore, higher attention was given to examining those modules.

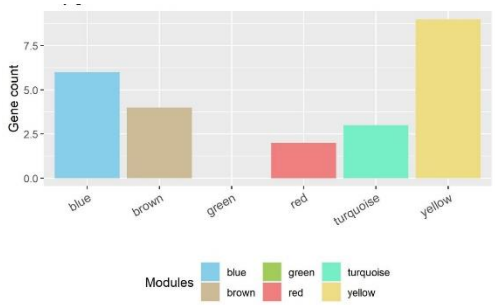

Figure 02: WGCNA modules identified for DEGs in *D. metel* transcriptome. The yellow module consisted of more DEGs while the red module for a smaller number of DEGs. No genes were observed in green module.

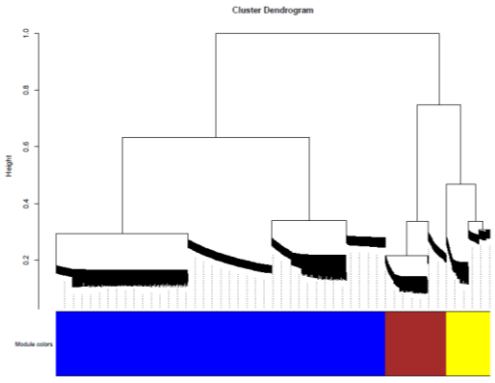

Figure 03: Known genes for withanolide biosynthesis pathway present in modules blue, brown, and yellow modules.

Table 01: Identified genes for further studies

Unigene	Gene Class	Symbol for CYP
TRINITY_DM8039_c0_g3_i1	Cytochrome P450	CYP83B1
TRINITY_DM483_c0_g1_i5	Cytochrome P450	CYP450 monooxygenase
TRINITY_DM7998_c0_g1_i1	Cytochrome P450	CYP87A2
TRINITY_DM472_c1_g1_i4	Cytochrome P450	CYP88A3
TRINITY_DM2774_c0_g1_i11	Cytochrome P450	CYP72A5
TRINITY_DM399_c0_g2_i2	Cytochrome P450	CYP97
TRINITY_DM4536_c0_g2_i2	Glycosyltransferase	UGT

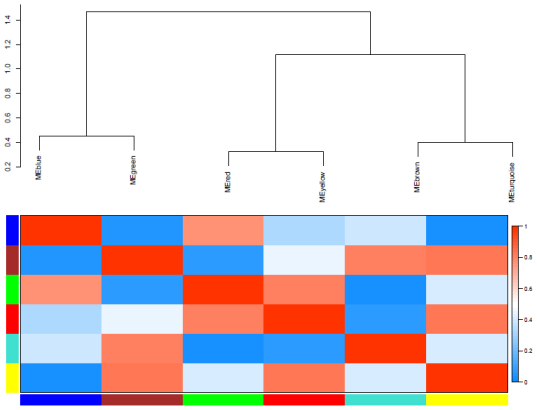

Figure 04: Correlation coefficient of modules with each other. Correlation 1 represents the higher module relationship while 0 represents the lower module relationship with each other.

Unigenes identified in WGCNA Unigenes co-expressed together or at least present in two co-expression methods were selected as the final possible candidate genes in the withanolide biosynthesis pathway for *the D. metel* transcriptome (Table 01).

4.0 CONCLUSION

Genes named DM483, DM472, DM2774, DM8039, DM7998, DM399, and DM4536 were identified as potential candidate genes for the withanolides biosynthesis pathway in *D. metel* leaf and flower tissues. The isolated genes were further characterized via *in silico* methods, phylogenetic trees, and multiple sequence alignment methods. This study shows the significance of using WGCNA analysis on identifying key modules and hub genes in Withanolides biosynthesis pathway in medicinal plants. However further *in silico* and *in vitro* characterizations are needed to carry on to confirm the genes functions and phenotypic expression in withanolides production in *D.metel* plants.

REFERENCES

Akhtar, T. A., Matsuba, Y., Schauvinhold, I., Yu, G., Lees, H. A., Klein, S. E., & Pichersky, E. (2013). The tomato cis–prenyltransferase gene family. *The Plant Journal*, *73*(4), 640-652.

Chaurasiya, N. D., Sangwan, N. S., Sabir, F., Misra, L., & Sangwan, R. S. (2012). Withanolide biosynthesis recruits both mevalonate and DOXP pathways of isoprenogenesis in Ashwagandha Withania somnifera L.(Dunal). *Plant cell reports*, *31*(10), 1889-1897.

Courtney A. Hollender, Chunying Kang, Omar Darwish, Aviva Geretz, Benjamin F. Matthews, Janet Slovin, Nadim Alkharouf, Zhongchi Liu, Floral Transcriptomes in Woodland Strawberry Uncover Developing Receptacle and Anther Gene Networks , *Plant Physiology*, Volume 165, Issue 3, July 2014, Pages 1062–1075, https://doi.org/10.1104/pp.114.237529

Crozier, A., Jaganath, I.B. and Clifford, M.N. (2007) Phenols, polyphenols and tannins: *Plant Secondary Metabolites: Occurrence, Structure and Role in the Human Diet.* Blackwell Publishing, pp. 1-24. ISBN 1405125098

De Bernonville, T. D., Papon, N., Clastre, M., O'Connor, S. E., & Courdavault, V. (2020).

Identifying missing biosynthesis enzymes of plant natural products. *Trends in Pharmacological Sciences*, *41*(3), 142-146.

Dhar, N., Razdan, S., Rana, S., Bhat, W. W., Vishwakarma, R., & Lattoo, S. K. (2015). A decade of molecular understanding of withanolide biosynthesis and in vitro studies in Withania somnifera(L.) dunal: Prospects and perspectives for pathway engineering. *Frontiers in Plant Science*, *6*(NOVEMBER), 1–20. https://doi.org/10.3389/FPLS.2015.01031/BIBTEX

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PloS one*, *6*(10), e26683.

Dugé de Bernonville, T., Carqueijeiro, I., Lanoue, A., Lafontaine, F., Sánchez Bel, P., Liesecke, F., ... & Courdavault, V. (2017). Folivory elicits a strong defense reaction in Catharanthus roseus: metabolomic and transcriptomic analyses reveal distinct local and systemic responses. *Scientific Reports*, *7*(1), 1-14.

Frey M, Chomet P, Glawischnig E, Stettner C, Grün S, Winklmair A, Eisenreich W, Bacher A, Meeley RB, Briggs SP, Simcox K, Gierl A. Analysis of a chemical plant defense mechanism in grasses. Science. 1997 Aug 1;277(5326):696-9. doi: 10.1126/science.277.5326.696. PMID: 9235894.

Gamboa-Tuz, S. D., Pereira-Santana, A., Zamora-Briseño, J. A., Castano, E., Espadas-Gil, F., Ayala-Sumuano, J. T., ... & Rodríguez-Zapata, L. C. (2018). Transcriptomics and co-expression networks reveal tissue-specific responses and regulatory hubs under mild and severe drought in papaya (Carica papaya L.). *Scientific reports*, *8*(1), 1-16.

Gillis J, Pavlidis P (2012) "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. PLOS Computational Biology 8(3): e1002444. https://doi.org/10.1371/journal.pcbi.1002444

Glotter, E. (1991). Withanolides and related ergostane-type steroids. *Natural product reports*, *8*(4), 415-440.

Gupta, P., Goel, R., Agarwal, A. V., Asif, M. H., Sangwan, N. S., Sangwan, R. S., & Trivedi, P. K. (2015). Comparative transcriptome analysis of different chemotypes elucidates withanolide biosynthesis pathway from medicinal plant Withania somnifera. *Scientific reports*, *5*(1), 1-13.

Gupta, P., Goel, R., Pathak, S., Srivastava, A., Singh, S. P., Sangwan, R. S., ... & Trivedi, P. K. (2013). De novo assembly, functional annotation and comparative analysis of Withania somnifera leaf and root transcriptomes to identify putative genes involved in the withanolides biosynthesis. *PLoS One*, *8*(5), e62714.

Higashi, Y., & Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. *Plant, cell & environment*, *36*(9), 1597-1606.

Huihui Gao, Yan Wang, Wei Li, Yongzhe Gu, Yongcai Lai, Yingdong Bi, Chaoying He, Transcriptomic comparison reveals genetic variation potentially underlying seed developmental evolution of soybeans, *Journal of Experimental Botany*, Volume 69, Issue 21, 12 October 2018, Pages 5089–5104, https://doi.org/10.1093/jxb/ery291

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., ... & Aharoni, A. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, *341*(6142), 175-179.

Jacobowitz, J. R., & Weng, J. K. (2020). Exploring uncharted territories of plant specialized metabolism in the postgenomic era. *Annual review of plant biology*, *71*(1).

Jia, D., Chen, C., Chen, C., Chen, F., Zhang, N., Yan, Z., & Lv, X. (2021). Breast cancer case identification based on deep learning and bioinformatics analysis. *Frontiers in Genetics*, *12*, 628136.

Ju Eun Jeon, Jung-Gun Kim, Curt R. Fischer, Niraj Mehta, Cosima Dufour-Schroif, Kimberly Wemmer, Mary Beth Mudgett, Elizabeth Sattely,A Pathogen-Responsive Gene Cluster for Highly Modified Fatty Acids in Tomato,Cell,180, 1,2020,176-

187.e19,00928674,https://doi.org/10.1016/j.cell.2019.11.037

Kazuki Saito, Masami Y. Hirai, Keiko Yonekura-Sakakibara,Decoding genes with coexpression networks and metabolomics – 'majority report by precogs',Trends in Plant Science,Volume 13, Issue 1,2008, 36-43,ISSN 1360-1385,https://doi.org/10.1016/j.tplants.2007.10.006.

Keiko Yonekura-Sakakibara, Takayuki Tohge, Fumio Matsuda, Ryo Nakabayashi, Hiromitsu Takayama, Rie Niida, Akiko Watanabe-Takahashi, Eri Inoue, Kazuki Saito, Comprehensive Flavonol Profiling and Transcriptome Coexpression Analysis Leading to Decoding Gene–Metabolite Correlations in *Arabidopsis* , *The Plant Cell*, Volume 20, Issue 8, August 2008, Pages 2160–2176, https://doi.org/10.1105/tpc.108.058040

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 1-13.

Li, X., Zhang, L., Ahammed, G. J., Li, Z. X., Wei, J. P., Shen, C., ... & Han, W. Y. (2017). Stimulation in primary and secondary metabolism by elevated carbon dioxide alters green tea quality in Camellia sinensis L. *Scientific reports*, *7*(1), 1-12.

Misico, R. I., Nicotra, V. E., Oberti, J. C., Barboza, G., Gil, R. R., & Burton, G. (2011). Withanolides and related steroids. *Progress in the Chemistry of Organic Natural Products Vol. 94*, 127-229.

Mutwil, M. (2020). Computational approaches to unravel the pathways and evolution of specialized metabolism. *Current Opinion in Plant Biology*, *55*, 38–46. https://doi.org/10.1016/J.PBI.2020.01.007

Nikolić, D., Gödecke, T., Chen, S. N., White, J., Lankin, D. C., Pauli, G. F., & van Breemen, R. B. (2012). Mass spectrometric dereplication of nitrogen-containing constituents of black cohosh (Cimicifuga racemosa L.). *Fitoterapia*, *83*(3), 441-460.

Nützmann, H.-W., Huang, A. and Osbourn, A. (2016), Plant metabolic clusters – from genetics to genomics. New Phytol, 211: 771-789. https://doi.org/10.1111/nph.13981

Piasecka, A., Jedrzejczak-Rey, N., & Bednarek, P. (2015). Secondary metabolites in plant innate immunity: conserved function of divergent chemicals. *New Phytologist*, *206*(3), 948-964.

Seca, A. M., & Pinto, D. C. (2019). Biological potential and medical use of secondary metabolites. *Medicines*, *6*(2), 66.

Senthil, K., Thirugnanasambantham, P., Oh, T. J., Kim, S. H., & Choi, H. K. (2015). Free radical scavenging activity and comparative metabolic profiling of in vitro cultured and field grown Withania somnifera roots. *PLoS One*, *10*(4), e0123360.

Serin, E. A., Nijveen, H., Hilhorst, H. W., & Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. *Frontiers in plant science*, *7*, 444.

Tai, Y., Liu, C., Yu, S., Yang, H., Sun, J., Guo, C., ... & Wan, X. (2018). Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). *Bmc Genomics*, *19*(1), 1-13.

Tissier, A. (2012). Glandular trichomes: what comes after expressed sequence tags? *The Plant Journal : For Cell and Molecular Biology*, *70*(1), 51–68. https://doi.org/10.1111/J.1365-313X.2012.04913.X

Tripathi, S., Jadaun, J. S., Chandra, M., & Sangwan, N. S. (2016). Medicinal plant transcriptomes: the new gateways for accelerated understanding of plant secondary metabolism. *Plant Genetic Resources*, *14*(4), 256-269.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., ... & Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, cell & environment*, *32*(12), 1633-1651.

Wang, X., Pan, Y. J., Chang, B. W., Hu, Y. B., Guo, X. R., & Tang, Z. H. (2016). Ethylene-Induced Vinblastine Accumulation Is Related to Activated Expression of Downstream TIA

Pathway Genes in Catharanthus roseus. *BioMed Research International*, *2016*. https://doi.org/10.1155/2016/3708187

Yang, Y., Qi, M., & Mei, C. (2004). Endogenous salicylic acid protects rice plants from oxidative damage caused by aging as well as biotic and abiotic stress. *The Plant Journal*, *40*(6), 909-919.

Zhang, H., Samadi, A. K., Cohen, M. S., & Timmermann, B. N. (2012), Antiproliferative withanolides from the solanaceae: A structure-activity study. *Pure and Applied Chemistry*, *84*(6), 1353–1367.

Zheng, J., He, C., Qin, Y., Lin, G., Park, W. D., Sun, M., ... & Liu, S. (2019). Co-expression analysis aids in the identification of genes in the cuticular wax pathway in maize. *The Plant Journal*, *97*(3), 530-542.