Identifying metabolic pathways of *Catharanthus roseus* seedlings affected by methyl-jasmonate and ethylene

Thierry Haddad

E-mail: thierry.haddad@wur.nl
Student ID: 911 031 296 110
Group: The Periwinkles

Abstract

Catharanthus roseus is a flower that produces two compounds used in the treatment of cancer. The composite MIA-pathway resulting in these compounds is not fully understood and key steps are missing. Various studies have indicated that the usage of methyl-jasmonate (MJ) and ethylene (ET) increases production of these two compounds. To further annotate and understand the effect of MT and ET, the group performed a pathway enrichment analysis on differentially expressed genes (DEG), where RNA-seq data was used from Pan *et al*. After preprocessing of the reads, the transcripts were aligned to the transcriptome of genome v2.0, as well as ran through Kallisto for normalized counts; TPM. Clustering was performed on the samples for potential outliers and MJ 1 was ommitted due to being mislabelled. BLASTx against SwissProt was used for transcript identification. The annotation and TPM were combined for a pathway enrichment analysis using KEGG. Results showed that various metabolic pathways were differentially expressed after MJ or ET treatment, including the MIA pathway. However, the strictness of the BLAST annotation had direct impact on how many MIA-related transcripts were found, where allowing more homologs per transcript greatly increased the amount of transcripts mapped in the MIA pathway. Our results differ from Pan *et al*., as we found that ET was a lot less potent in inducing DEGs than they have found.

1. Introduction

Vincristine and vinblastine are two terpenoid indole alkaloid (TIA) compounds that are being applied in chemotherapy for the treatment of several types of cancer. These compounds are the products of a complex pathway, the Monoterpene Indole Alkaloids(MIA)-pathway in the species *Catharanthus roseus*, which is not fully understood at the time of writing as various steps in the pathway are not annotated. In summary, the protein enters the MIA-pathway through the Terpenoid Back synthesis where it will undergo various biochemical changes, before ending up in the Indole Alkaloid biosynthesis pathway. In the Indole Alkaloid pathway, the compound can traverse a variety of paths, but it is speculated that 4,21-dehydro-geissoschizine converts to stemmadenine through an as of yet unclear reaction. From there it can go through either the Vindoline or Catharanthine path to converge in to 3',4'-anhydro-vinblastine, the precursor to vinblastine and vincristine.

A previous study has found evidence for an increase in pathway-related transcripts expression after having the seedlings of *C. roseus* induced in either methyl-jasmonate (MJ) or ethylene (ET)[1]. In this study, we tried our own approach towards the identification of pathways that are affected by having *C. roseus* samples induced with either MJ or ET, making use of the same RNA-seq dataset and transcriptome as used by Pan *et al*. Our study was predominantly an exploratory one, where we would test our own version of a differential expression pipeline and compare it to the previosuly mentioned study. However, our approach unfolded differently than expected, as we discovered several errors in the data and an opposing conclusion.

2. Design

Note: The various Python and R scripts noted below can be found on the Wageningen University & Research (WUR) Altschul server at /local/data/course/project/groups/periwinkles/FINAL.

2.1 RNA-seq reads and quality control

This study made use of the RNA-seq reads found on the WUR Altschul server under /local/data/course/project/RNAseq/SRP095740, which are the reads used by Pan *et al.*[1] Additionally, these reads are stored on the Sequence Read Archive as accession number SRP095740 and on NCBI BioProject as PRJNA358259. The reads used consists of nine samples over three conditions: control (CK1, CK2, CK3), MJ-induced (MJ1, MJ2, MJ3) and ET-induced (ET1, ET2, ET3). Due to an anomaly in the provided data, only the forward reads were used instead of the paired-end reads and sample CK3 was ommitted. The anomaly is described in the results. The total amount of used reads amount up to 182.730.648.

Quality control of the RNA-seq reads was performed with the commandline version of FASTQC v0.11.4 on the WUR Altschul server. For this, script "unzipping_trimming.py" was used, which will unzip the FASTQ-based RNA-seq reads and perform trimming with fastq_quality_trimmer using a threshold of 33 and a ASCII quality offset of 33 (Illumina HiSeq 2000).

2.2 Mapping with Bowtie2

As the mapping is performed against a transcriptome instead of a genome file and intron-exon junctions are not present, we opted for the use of Bowtie2 as recommended in the "RNA-seq best-practice survey"[2]. Therefore, indexing of the reference transcriptome and mapping of the trimmed reads is performed with the commandline version of Bowtie2 v2.2.6. The "—no-unal" argument was added to reduce BLAST and annotation times for reads that failed to align. The reference transcriptome used as Bowtie2 index is the transcriptome of genome V2 assembled by Franke *et al.*[3]. Initially, the assembled transcriptome from Pan *et al.* was used, but this resulted in very poor alignment rates and after consultation the Franke *et al.* transcriptome was used. The index was created with bowtie2-build. The output was redirected to a SAM file and SamTools v0.1.19-96b5f2294a was used for the conversion of SAM to BAM for the downstream annotation.

2.3 Mapping quantification with Kallisto

A pseudoalignment for the efficient mapping of reads and the quantification and normalization of readcounts was created with Kallisto 0.44.0 as recommended by the RNA-seq survey[2]. "kallisto index" Was used for the creation of an index that is required for the pseudoalignment, based on the earlier mentioned transriptome of genome V2 from Franke *et al.*[3]. The trimmed, forward reads were used for mapping. "kallisto quant" Was used for the quantification of mapped reads. Arguments "—single", "-l 90" and "-s 9" were added to indicate the quantification is for single reads only, the fragment length is 90 and the expected standard deviation is 9 respectively. The discussion dives further into these values. The resulting transcripts per million (TPM) values were parsed into a tab-delimited format for further usage in DESeq2.

2.4 Clustering and identifying differentially expressed genes with DESeq2

The clustering of samples and identification of differentially expressed genes (DEG) was performed by the R BioConductor package DESeq2. All clustering and expression analysis (including downstream annotation) was done on a control-to-condition basis (CK-ET, CK-MJ) as our objective is predominantly finding differences between ET/MJ induced samples versus the control groups, rather than the difference between ET and MJ and control.

Clustering was done using the Euclidian distance metric and complete-linkage hierarchical clustering approach with the "hclust" method on log2-transformed values of the size-factor-normalized TPM values to account for sequencing depth differences.

Gene expression dispersion was calculated with R's "estimateDispersions". The DEGs were calculated using R's "nbinomWaldTest" method. For the MA-plot, the log2-normalized fold change values of the DEGs were used. The heatmap of clustered control genes, that were selected based on literature[7], was made using the "heatmap" method and the principal component analysis was created with the "plotPCA" method. All of the above steps are performed by the script "Differential.r". However, this script currently does not work on the WUR Altschul server due to having a different R version and therefore not being able to install the required libraries.

2.5 Annotating transcripts with BLASTx

Mapped Bowtie2 transcripts in SAM format were parsed by "blast.py" and annotated using NCBI's BLASTx engine. Mostly default parameters were used, including the organism, as we aimed to also find very close homologs that could potentially provide valueable insight on missing annotations in the pathway. However, for the reference database we have chosen to only use SwissProt, as we preferred very well annotated proteins (even if from a different species, but a very close homolog) over potentially non-verified proteins (e.g. nr/nt database) as these could add more noise to the pathway annotation rather than clarify or enhance our understanding of it. For the cut-off e-value we selected a standard error margin of 5% with 5.0×10^{-2}. From the results, the best hit per transcript was selected and parsed to a comma-separated-values file, which was in turn parsed by "blastx_parser.py" for KEGG annotation. Additionally, this step was repeated with all hits above the e-value cut-off, instead of only the best hit.

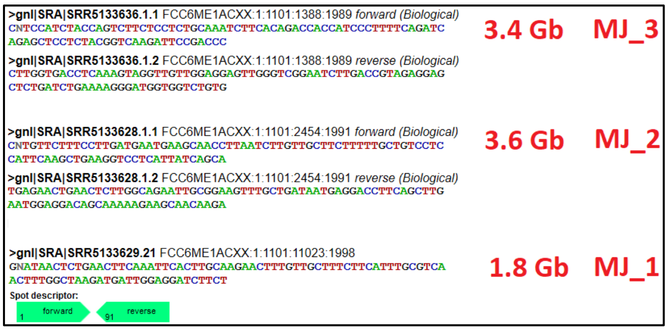
2.6 Pathway annotation using KEGG

The parsed annotated transcripts for CK-MJ and CK-ET gotten from BLASTx were used with UniProt's Retrieve/ID Mapping tool on https://www.uniprot.org/uploadlists/, where UniProtKB AC/IDs were converted to KO (KEGG Ontology) accession numbers for the identification of differentially expressed pathways. The resulting IDs were used for KEGG's Search&Color Pathway Mapper on https://www.genome.jp/kegg/tool/map_pathway2.html with all parameters set to default.

3. Results

3.1 Read quality and alignment rates

As described in section 2.1, we've discovered an anomaly in the provided RNA-seq dataset, including on the Sequence Read Archive. Where Pan *et al*. indicated that all nine samples were paired-end reads, closer inspection has shown that in fact set MJ1 was missing the reverse part of the reads, as can be seen in Figure 1. Where MJ2 and MJ3 both have forward and reverse reads and are >3 Gb in size, MJ1 has roughly half the size and is missing the reverse (and all reads are 90 nc long instead of 180). Supplementary file Table S1 in the paper of Pan *et al*. shows that MJ1 should be roughly 3.6 Gb in size. This has lead us to conclude Pan *et al*. have made an error in providing the correct data and therefore we opted for only forward reads as we preferred having more biological replicates over paired-end reads as we hypothesized that for the identification of DEGs, a more trustworthy average per condition type (due to biological variation) would be more valueable than potentially better mapping.

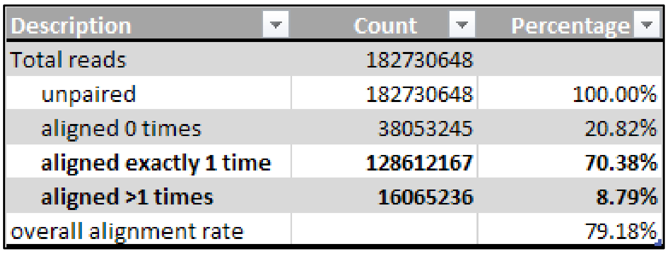


```
>gnl|SRA|SRR5133636.1.1 FCC6ME1ACXX:1:1101:1388:1989 forward (Biological)
CNTCCATCTACCAGTCTTCTCCTCTGCAAATCTTCACAGACCACCATCCCTTTTCAGATC
AGAGCTCCTCTACGGTCAAGATTCCGACCC
>gnl|SRA|SRR5133636.1.2 FCC6ME1ACXX:1:1101:1388:1989 reverse (Biological)
CTTGGTGACCTCAAAGTAGGTTGTTGGAGGAGTTGGGTCGGAATCTTGACCGTAGAGGAG
CTCTGATCTGAAAAGGGATGGTGGTCTGTG


>gnl|SRA|SRR5133628.1.1 FCC6ME1ACXX:1:1101:2454:1991 forward (Biological)
CNTGTTCTTTCCTTGATGAATGAAGCAACCTTAATCTTGTTGCTTCTTTTTGCTGTCCTC
CATTCAAGCTGAAGGTCCTCATTATCAGCA
>gnl|SRA|SRR5133628.1.2 FCC6ME1ACXX:1:1101:2454:1991 reverse (Biological)
TGAGAACTGAACTCTTGGCAGAATTGCGGAAGTTTGCTGATAATGAGGACCTTCAGCTTG
AATGGAGGACAGCAAAAAGAAGCAACAAGA


>gnl|SRA|SRR5133629.21 FCC6ME1ACXX:1:1101:11023:1998
GNATAACTCTGAACTTCAAATTCACTTGCAAGAACTTTGTTGCTTTCTTCATTTGCGTCA
ACTTTGGCTAAGATGATTGGAGGATCTTCT
```

Figure 1: Comparison of the three methyl-jasmonate induced samples RNA-seq reads. Notice how MJ_1 has half the bases of MJ_2 and MJ_3 and the reverse strand appears to be missing.

Trimming the reads with fastq_quality_trimmer resulted in a minor increase in read quality according to FASTQC, as both before and after no adapter sequences were found and afterwards some reads were trimmed ~2-5 nucleotides.

Initial mapping with Bowtie2 was performed using Pan *et al*.'s transcriptome, but this resulted in very poor alignment rates in the range of 17-20%, even if using very sensitive local search settings. After consultation we decided to further our research with the more well-curated transcriptome of genome V2 made by Franke *et al*. This resulted in a very significant increase in alignment rate, reaching approximately 79% (Figure 2). A similar result was found when mapping with Kallisto, resulting in an alignment rate of approximately 82% for the pseudoalignment.



Description	Count	Percentage
Total reads	182730648	
unpaired	182730648	100.00%
aligned 0 times	38053245	20.82%
aligned exactly 1 time	**128612167**	**70.38%**
aligned >1 times	**16065236**	**8.79%**
overall alignment rate		79.18%

Figure 2: Bowtie2 Alignment overview of forward reads of all samples (CK, MJ, ET) against the genome V2 transcriptome.

3.2 Clustering the samples

Clustering was performed with R BioConductor's package DESeq2 and generated some interesting results in both the Euclidean clustering and the principal component analysis (PCA).

Starting with the PCA, in an ideal hypothetical situation one would expect the largest variance to be between the control samples and the ET/MJ induced samples. However, the PCA notified us of another anomaly in the dataset: control sample CK3 had much more variance with the other control samples than with the ET induced samples (Appendix Figure 6). This was supported by the fact that after we reran the PCA and downstream DEG analysis without control sample CK3 we able to find many more DEGs. This leads us to believe that there either a laboratory error has been made or, more likely, Pan *et al*. mislabelled their sample and it should have been labelled as an ET sample instead. However, this remains a hypothesis and in order to avoid downstream errors due to wrong assumptions, we continued our research without control sample CK3.

The generated heatmap (Figure 3) based on control genes did not display what we expected, as expression levels are not always in line with literature. An example is ORCA3, which should be strongly up-regulated in MJ, somewhat in ET and not in CK [7]. Here, only one MJ sample and one ET sample are up-regulated, while two ET samples are effectively down-regulated. We've briefly hypothesized that this might be due to taking the samples at different locations with some time intervals between and amplified by a general lack of biological replicates.

Euclidean clustering provided us with another interesting finding, in that ethylene sample ET3 clustered more closely with control samples CK1 and CK2 (Figure 4). This issue persisted when using a different method (correlation instead of Euclidean) and linkage (single instead of complete). However, the distance between ET3 and CK1 and CK2 remained considerable and due to a clear PCA and time constraints, we opted to keep ET3 as is (but it remains an interest for further research).

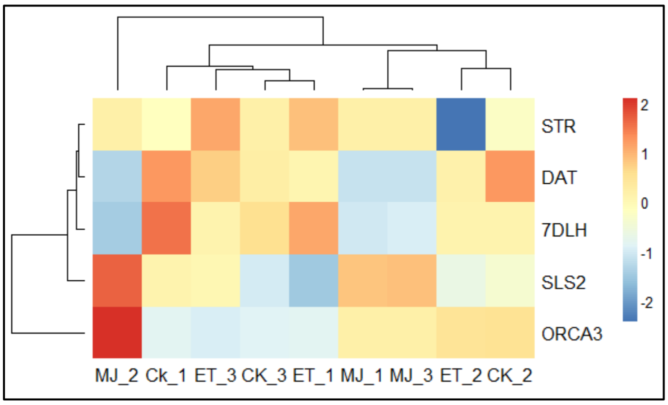


Figure 3: Heatmap of several control genes clustered on RNA-seq expression levels, where blue indicates under-expression and red over-expression. Horizontally the nine individual samples are listed, vertically the control genes.

3.3 Differentially expressed genes

Appendix Figure 7 shows various plots created by the expression analysis using DESeq2 for control against methyl-jasmonate and control against ethylene respectively.

7.A displays the volcano-plots of MJ and ET, where the x-axis shows the log2-foldchange values and the y-axis the log10-p-values. The striped green line displays the standard p-value cut-off of 0.05. Each dot represents one gene and red-coloured dots represent significantly (p-value < 0.05) differentially expressed genes.

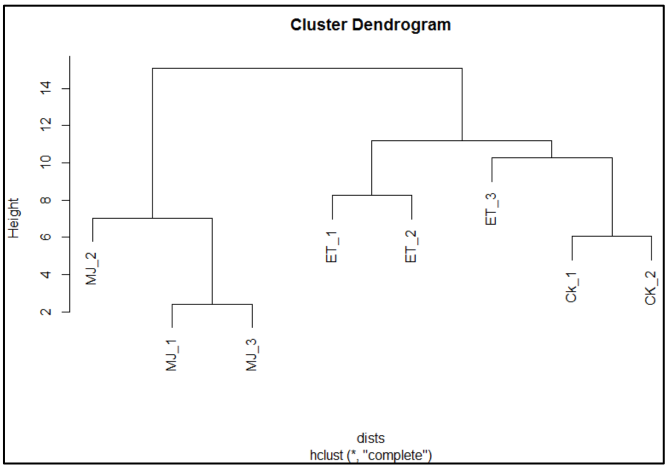


Figure 4: Euclidean distance clustering with complete linkage in R of all samples, after omitting CK3. ET_3 is placed in a clade with CK_1 and CK_2 rather than with the other ET samples, but the distance appears to be large.

Notice how both MJ and ET find relatively few down-regulated genes (ET slightly more than MJ) compared to up-regulated genes. Especially MJ seems to find many DEGs with very low p-values. This seems to contradict the findings by Pan *et al.* where they found ET to have a more significant effect on DEGs than MJ.

7.B are the PCAs for CK-MJ and CK-ET respectively. Especially the MJ PCA shows a very clear distinction between the CK and MJ samples, where roughly 88% of the variance can be explained by the first principal component (PC1, x-axis), indicating a large and distinctive difference between CK and MJ. PC2 explains roughly 9% of the variance and shows the in-condition variance. However, the PCA for CK-ET shows considerably less explained variance in PC1 at roughly 53% and a larger PC2 (at 31%) than CK-MJ. These findings support the number of DEGs indicated in the volcano-plots, as the CK-MJ samples seem to be much more distinct than CK-ET, resulting in more significant DEGs for CK-MJ.

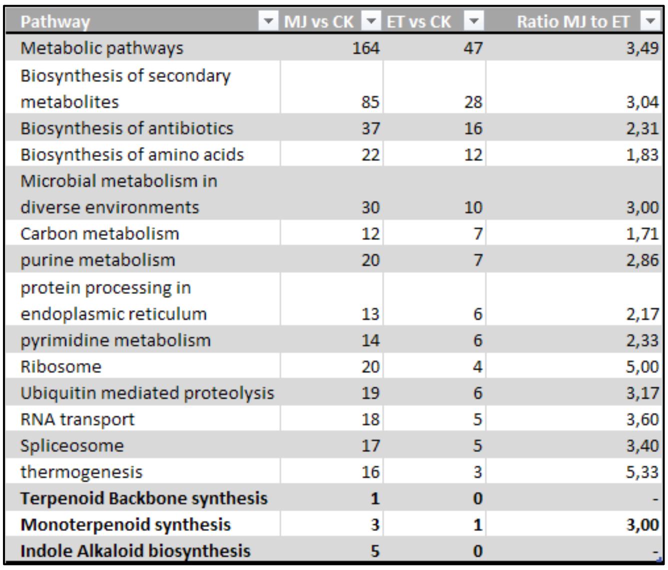
7.C shows the dispersion plots for CK-MJ and CK-ET where the mean of the normalized counts (x-axis) is laid out against the dispersion (y-axis) which indicates the variance divided by the squared mean to give an overview of how the variance is spread over the read counts. Higher mean read counts tend to have less dispersion. The CK-ET plot indicates it has several genes with high dispersion (and also variance if looking at higher mean read count numbers) which might also support our different finding of CK-ET having less DEGs than CK-MJ.

7.D displays the MA-plots for both CK-MJ and CK-ET, where the x-axis is the mean of normalized RNA-seq counts and the y-axis the log2 fold change. Each plot compares, per gene, how it is expressed in group 1 (CK) and group 2 (MJ or ET). If the gene behaves similar in both group 1 and 2, it will be placed along the 0 log2 fold change as no change is measured. Red dots indicate genes that have a significant difference in expression between the groups, where a higher mean of normalized counts allows for lower log2 fold change for it to be significant. Like in previous findings, inducing *C. roseus* with MJ seems to generate a lot more DEGs than inducing with ET.

3.4 BLAST and annotation with KEGG pathways

BLASTx-ing against the SwissProt database and conversion of UniProt Accession numbers to KO IDs for the usage in KEGG was performed for both the CK-MJ and CK-ET samples. This resulted in 634 and 205 KO IDs respectively. Both lists were parsed in the KEGG Search&Colour Pathway Mapper, where CK-MJ found a total of 1536 pathway hits and CK-ET 487. The difference in numbers is expected, given the larger amount of DEGs found in CK-MJ than in CK-ET. A brief display of the top pathway hits can be seen in Figure 5. A lot of general house-keeping pathways are differentially expressed in both CK-ET and CK-

MJ, however more-so in CK-MJ (often a factor 2 to 3). In bold are the pathways of interest as they participate in the production of vincristine and vinblastine. MJ-induced samples appear to significantly differentially express one Terpenoid Backbone synthesis gene, three Monoterpenoid synthesis genes and five Indole Alkaloid biosynthesis genes. ET-induced samples, on the other hand, only significantly affect one gene in the Monoterpenoid synthesis (Appendix Figure 8). This leads us to believe that methyl-jasmonate has a much more profound effect on the potential production of vincristine and vinblastine than ethylene, which is not in line with the findings of Pan *et al.* as they have found the opposite effect.



Pathway	MJ vs CK	ET vs CK	Ratio MJ to ET
Metabolic pathways	164	47	3,49
Biosynthesis of secondary metabolites	85	28	3,04
Biosynthesis of antibiotics	37	16	2,31
Biosynthesis of amino acids	22	12	1,83
Microbial metabolism in diverse environments	30	10	3,00
Carbon metabolism	12	7	1,71
purine metabolism	20	7	2,86
protein processing in endoplasmic reticulum	13	6	2,17
pyrimidine metabolism	14	6	2,33
Ribosome	20	4	5,00
Ubiquitin mediated proteolysis	19	6	3,17
RNA transport	18	5	3,60
Spliceosome	17	5	3,40
thermogenesis	16	3	5,33
Terpenoid Backbone synthesis	**1**	**0**	**-**
Monoterpenoid synthesis	**3**	**1**	**3,00**
Indole Alkaloid biosynthesis	**5**	**0**	**-**

Figure 5: Top 14 KEGG pathway results with additionally the Terpenoid Backbone synthesis, the Monoterpenoid synthesis and the Indole Alkaloid biosynthesis. There appears to be a factor 2 to 3 difference between MJ and ET.

Additionally, annotating with BLASTx and mapping to KEGG pathways was repeated for all potential homologs per transcript, instead of just the top hit (Appendix Figure 9). For MJ, this resulted in 8577 pathway hits, of which 13 Indole alkaloid biosynthesis and 14 Monoterpenoid biosynthesis. ET resulted in 3726 pathway hits, of which 6 Terpenoid backbone biosynthesis, 2 Monoterpenoid biosynthesis and 3 Indole alkaloid biosynthesis.

4. Discussion

Here our findings, the issues we faced and potential improvements to our study will be briefly discussed. As mentioned earlier, we have discovered several anomalies in the provided RNA-seq data from the Pan *et al*. study. Firstly, one of the methyl-jasmonate samples was devoid of the reverse reads and thus contained only half the bases of the other samples, leading to us only using the forward reads of the data set as we preferred more biological replicates over more secure paired-end reads. Second, at least one of the samples seems to have a worrisome variance with the other samples in that group, in this case the control group. One of the ethylene-induced groups, ET3, was also found to be clustering with the control group instead of the other ET samples. Given the sequence statistics given by the study, even though the SRA library also lacked the data, we don't expect the study to have incorrectly used the RNA-seq data. However, we do expect for them to potentially have mislabelled CK3 as a control group sample, while it shows a lot more similarity with the ET group. ET3 might also have been mislabelled, and in reality be a CK group sample, but the difference appears to be smaller than the first example and we were not able to test this hypothesis, resulting in us not taking this further into consideration. Further research might indicate whether this hypothesis holds true and whether the study from Pan *et al*. remains correct. Additionally, if the SRA library is updated with the correct paired-end data set, this research might be repeated to gauge whether our findings still hold.

Expanding further upon the reads used for the alignment, an additional note is that our usage of kallisto could be improved a bit. While we don't expect a very different result, two parameters were filled in to the best of our abilities, but might not be fully correct. These parameters were the fragment length and the standard deviation that are required to use when making use of single-end reads[6]. Due to time constraints and the usage of single-end reads, we were not able to find the used fragment length or calculate the true standard deviation. Therefore we opted for our base read-length, 90 and estimated a 10% standard deviation, 9. However, as the pseudoalignment of Kallisto had a very similar alignment rate as Bowtie2, we think these estimated parameters were very reasonable and don't expect the differences resulting from rectification to be more than minor.

Which leads into the DEG annotation with BLASTx. SwissProt was used as a reference for more well-curated and established protein annotations. While this, on first glance, might make the most sense given the situation of a not fully-understood pathway, it also leaves the question on whether there might have been related genes in a different reference database such as nr/nt, where less curation occurs. This would result in a trade-off between less, but more well-defined data and an over-abundance of data that might contain a lot of noise. The latter might prove fruitful if the ability for time-consuming annotation is present, but as we were limited due to time constraints, the SwissProt method seemed most feasible.

Lastly, a retrospect on the identification of pathways to which the identified DEGs belong. One first note to make is that after the conversion of UniProtKB IDs to KO IDs and putting these in the KEGG pathway tool, not all of the KO numbers were able to be linked to a pathway. This emphasizes the notion that our understanding of molecular pathways and annotation of such findings still remains very much a work-in-progress. Also, even when taking the difference in the amount of DEGs between MJ and ET in account, it's remarkable that ET found so few MIA-pathway related genes compared to MJ. It's also interesting to note that the different BLAST results generated a large difference in KEGG pathway hits, where the result that allowed for all homologue hits per transcript resulted in an expansive coverage of the pathways. This makes us believe there might be two possible explanations: 1) we've already gathered some annotation for the missing or poorly understood genes, but they are not properly linked to the pathway yet, or 2) the intermediate steps in the pathways might be caused by genes that share enough homology to be present in the BLASTx results. This could be an insight for further research on the pathways.

5. Conclusion

In this study, we've successfully substracted pathways with differentially expressed genes after they had been induced with MJ or ET. During the data pre-processing, various anomalies were found in the dataset provided by the study from Pan *et al.*, making their findings with the currently provided data irreproducible. We've worked around these anomalies to the best of our knowledge and produced a set of DEGs which yielded different results than previously mentioned study. The main difference is their finding of ET having a stronger effect than MJ for MIA-pathway expression, while we found the complete opposite effect. We've listed several shortcomings and potential improvements to our study and shown that even published studies are still capable of containing errors or inaccuracies.

6. Author contributions

- Found erroneous data on server and SRA;
- Map_reads_forward.py: Creating an index with Bowtie2, mapping reads to index and saving .sam;
- Aided with the unzipping and trimming script;
- Performed several of the DESeq2 plot/graph interpretations;
- Initial BLASTx, UniProt conversion and KEGG results.

References

[1] Pan, Y. J., Lin, Y. C., Yu, B. F., Zu, Y. G., Yu, F., & Tang, Z. H. (2018). Transcriptomics comparison reveals the diversity of ethylene and methyl-jasmonate in roles of TIA metabolism in Catharanthus roseus. *BMC genomics*, *19*(1), 508.

[2] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*(1), 13.

[3] Franke, J., Kim, J., Hamilton, J. P., Zhao, D., Pham, G. M., Wiegert-Rininger, K., ... & Buell, C. R. (2018). Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis. *ChemBioChem*.

[4] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

[5] Vom Endt, D., e Silva, M. S., Kijne, J. W., Pasquali, G., & Memelink, J. (2007). Identification of a bipartite jasmonate-responsive promoter element in the Catharanthus roseus ORCA3 transcription factor gene that interacts specifically with AT-Hook DNA-binding proteins. *Plant physiology*, *144*(3), 1680-1689.

[6] Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525.

[7] Zhang, H., Hedhili, S., Montiel, G., Zhang, Y., Chatel, G., Pré, M., ... & Memelink, J. (2011). The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *The Plant Journal*, *67*(1), 61-71.

Appendix

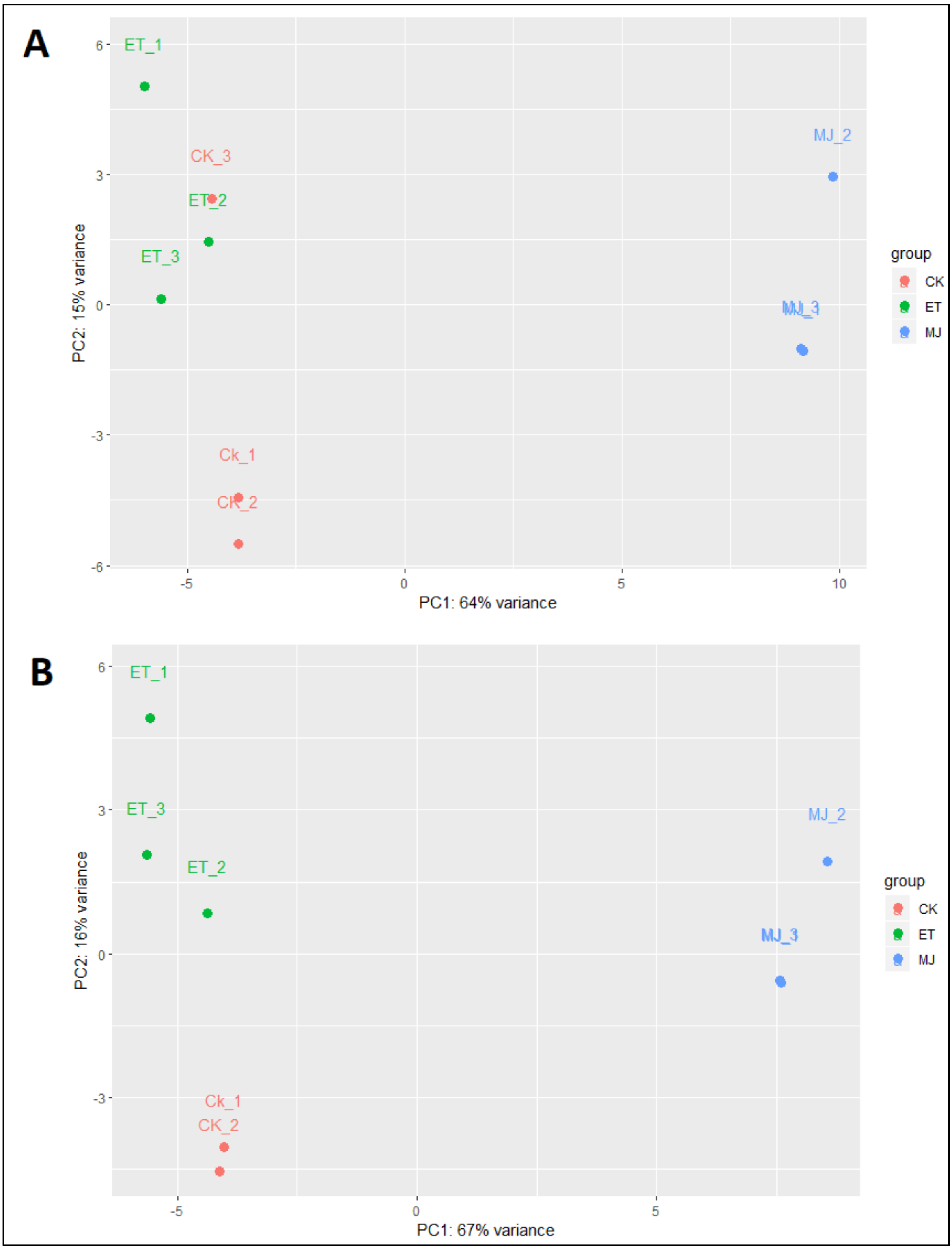


Figure 6: Principal component analysis before(A) and after(B) the removal of control sample CK3. Notice the distinct grouping of samples in section B. In section A it is clear that CK3 shows so much variance with the other CKs, that it cannot be reasonably expected to be valid data anymore. A mislabelling of an ET sample is not unthinkable.

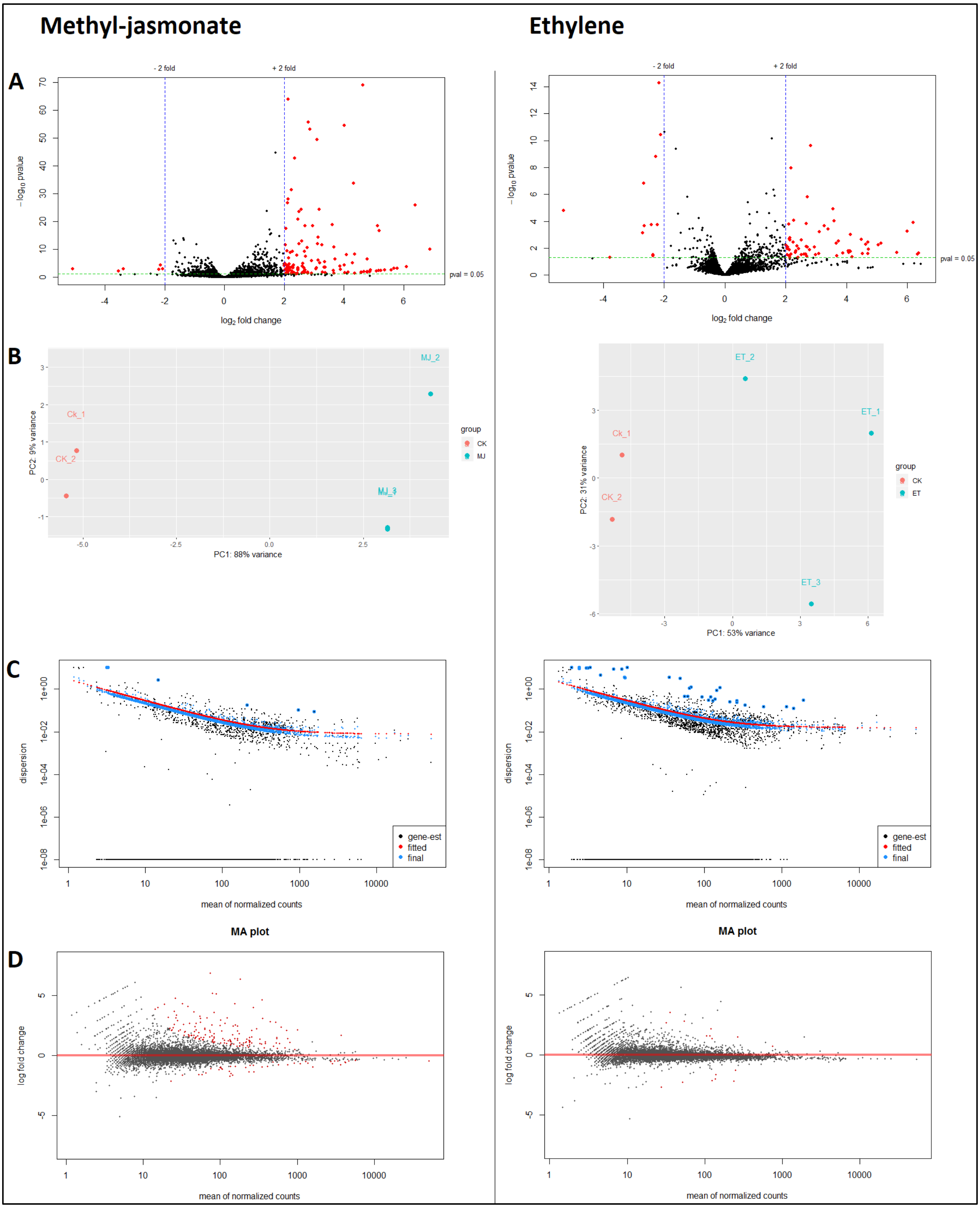


Figure 7: Various plots generated by DESeq2 in R for both the MJ-induced and ET-induced samples. A) Volcano-plot showing the fold change against significance, B) CK-MJ and CK-ET principal component analysis instead of an all-conditions PCA like before, C) dispersion plots laying the dispersion agains the mean of the normalized read counts, D) MA-plot displaying the log2 fold change of genes agains the mean of the normalized read counts, where red indicates significance.

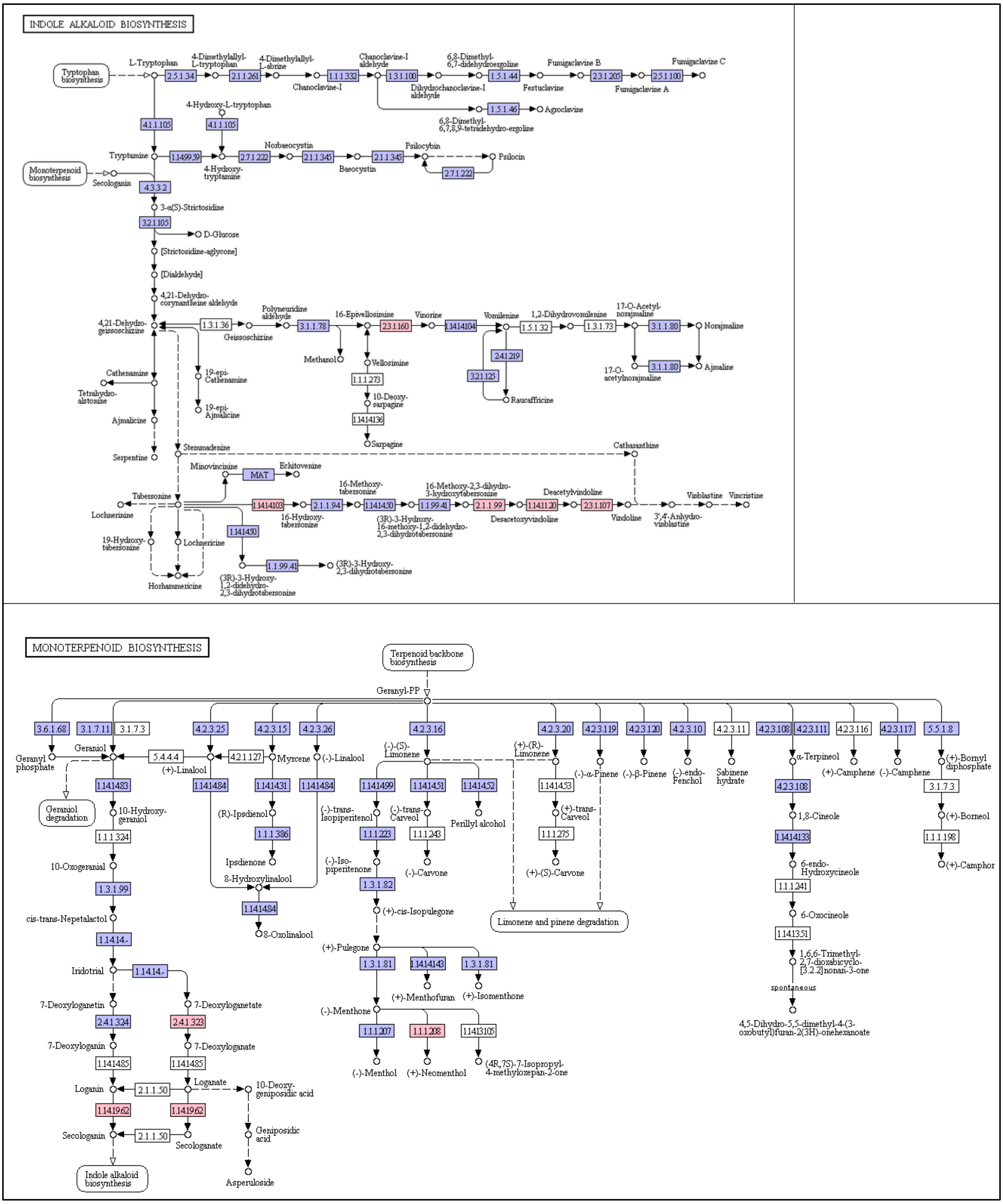


Figure 8: KEGG Indole Alkaloid biosynthesis and Monoterpenoid pathways for DEGs resulting from MJ-induced samples using only the top BLASTx hit per transcript. The DEGs can be identified by the red boxes found at predominantly the latter areas of the pathways.

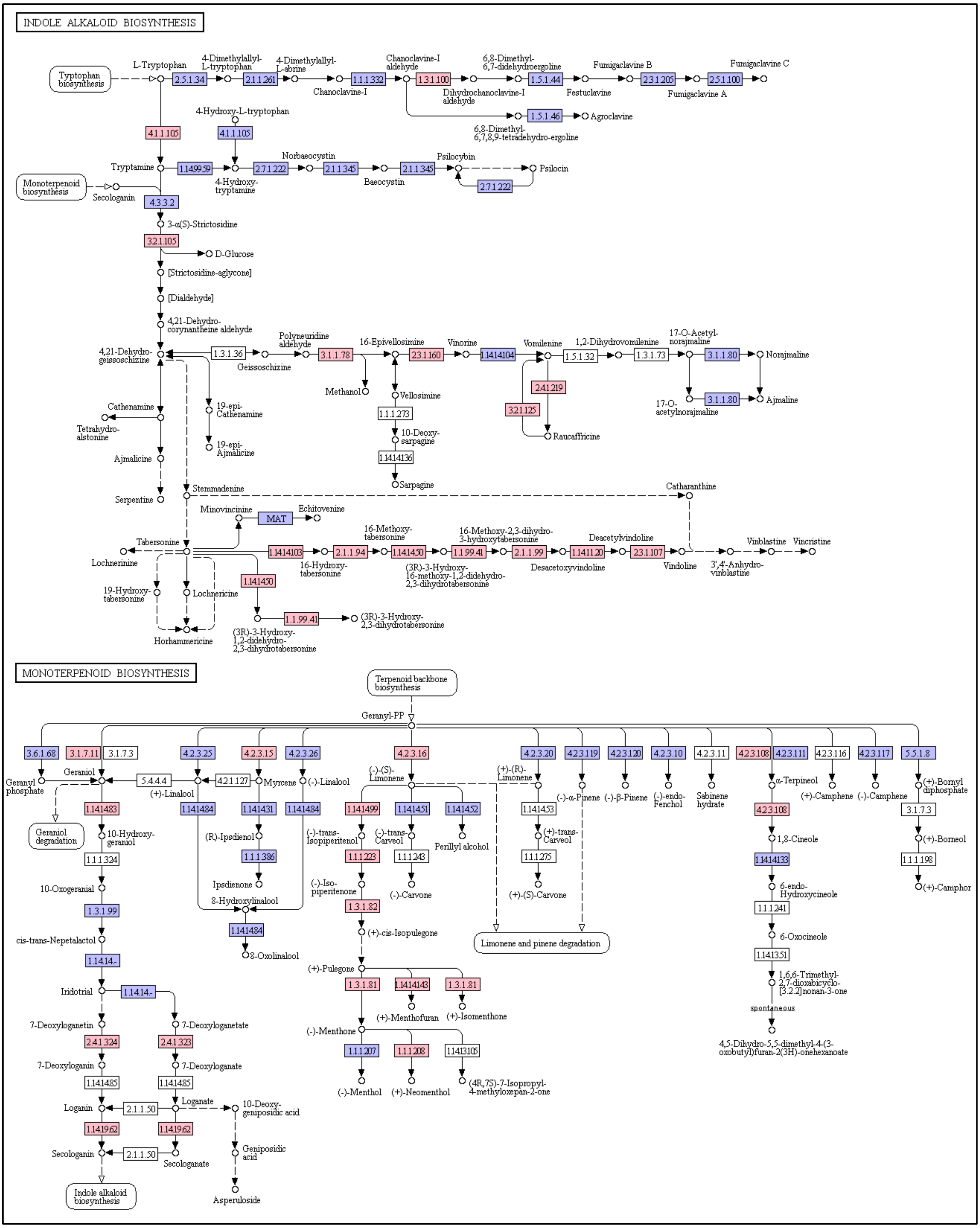


Figure 9: KEGG Indole Alkaloid biosynthesis and Monoterpenoid pathways for DEGs resulting from MJ-induced samples using all the BLASTx hits per transcript. The DEGs can be identified by the red boxes. Note how, compared to Figure 8, both pathways show a lot of linked DEGs.