Comparison of leaf and stem tissues in *C. canephora*

Vincent von Häfen and Flora Schlüter

Abstract

This study tries to clarify the functional differences of stem and leaf tissue for the species *Coffea canephora*. To this end a RNA-seq analysis was performed to find differentially expressed genes. In this study RNA-seq analysis is defined to include evaluating the quality of the raw RNA-seq data, mapping the accession data to the genome of *c. canephora*, performing differential expression analysis, a statistical evaluation as well as functional enrichment analysis. The resulting gene ontology annotations from the functional enrichment analysis in the stem tissue related to the cytoskeleton, which aligns with the biological role of the stem in plant stability and transport. Similarly, the significantly enriched GO annotations of the leaf tissue related to photosynthesis, which again aligns with the biological role of the leaf as the primary site of photosynthesis in plants. These results confirm the existing knowledge of the scientific community, supporting the work of many other researchers.

Introduction

RNA-seq analysis is used to analyze which genes are being expressed in a cell, meaning which genes are being transcribed to messenger RNA. As opposed to DNA analysis, which should be nearly identical in all cells originating from the same organism, RNA-seq analysis can be used to gain insight into which proteins are being manufactured by a cell in a specific environment. It not only yields information about the proteins which are required for a cell to fulfill its functions, and therefore can be used to identify cell type, but also captures differences due to extracellular signals or other environmental stimuli.

This study compares the RNA-seq data originating in leaf and stem tissues from *Coffea canephora*, commonly known as Robusta coffee. *C. canephora* is a species of coffee that is grown extensively in West and Central Africa, Southeast Asia, and Brazil and one of the most widely traded commodities in the world. *C. canephora* is known for its strong flavor and high caffeine content, and is valued for its resilience to pests and diseases. It has the ability to grow in a wide range of environmental conditions and has a ploidy level of 2 with a DNA amount of 0.72C (pg)[1].

RNA-seq data was chosen for this study since it holds information about the expressed genes in the leaf and stem tissues and thus enables analysis of the tissue differences in mRNA expression and abundance. The aim of this study was to identify genes which show statistically significant differences in mRNA expression and abundance between the analyzed tissues and draw conclusions on functions of the two tissues based on these differences.

In this study RNA-seq analysis is defined to include evaluating the quality of the raw RNA-seq data, mapping the data to the genome of *c. canephora*, performing differential expression analysis and a statistical evaluation as well as functional enrichment analysis.

Materials and Methods

Data selection

The SRA accessions were downloaded in FASTQ format from the NCBI SRA database and originate from the study “Metabolite Profiling and Transcriptome Analysis Revealed the Conserved Transcriptional Regulation Mechanism of Caffeine Biosynthesis in Tea and Coffee Plants”, published by the journal of agriculture and food chemistry in 2022[2].

From the available RNA-seq data, two stem tissue samples as well as two leaf tissue samples of a *c. canephora*[3] plant were chosen. Each tissue sample additionally had one biological replicate, which was also included in order to reach a total of 8 separate samples.

Quality Evaluation

The data was specifically downloaded in FASTQ format because FASTQ files include quality information that was necessary to perform a quality evaluation using the software package FastQC[4]. FastQC tests for several different aspects to determine the quality of the given data, namely: per base sequence quality, per base sequence quality score, per base sequence content, per base GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and kmer content.

[1] Noirot M, *et al* 2003. Genome size variations in diploid African Coffea species. Annals of Botany 92: 709-714.
[2] Zhang, Yanrui & Fu et al,.” Metabolite Profiling and Transcriptome Analysis Revealed the Conserved Transcriptional Regulation Mechanism of Caffeine Biosynthesis in Tea and Coffee Plants”, 2022, Journal of Agricultural and Food Chemistry”. 70. 10.1021/acs.jafc.1c06886.
[3] https://www.ncbi.nlm.nih.gov/bioproject/PRJNA798825
[4] “Trim Galore!” n.d. Babraham Bioinformatics. Accessed April 29, 2023. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Based on the FastQC reports, each sequence was trimmed with the software package Trimmomatic (Trimmomatic-0.39)[5].

The following specific parameters were used:

Parameter	Function
SE	uses the single-ended adapter file
ILLUMINACLIP:TruSeq3-SE.fa:2:30:10	Removes adapters and other illumina-specific sequences
HEADCROP:5	Cut 5 bases from the start of every read
SLIDINGWINDOW:4:15	Scans the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15
MINLEN:70	Drops reads below the 70 bases long
-phred33	Quality scores are Phred-33

Table 1.

Mapping to the Reference Genome

In order to map to a reference genome the genome of *C. canephora*[6] was downloaded from ensemble in the form of a cdna.all.fa file and a kallisto[7] index was built.

The trimmed SRA accessions were then pseudomapped to the indexed genome with the function kallisto quant using the following parameters: --single -l 180 -s 20

The result of the pseudomapping contains information on the count and abundance of the mapped reads, which are used for further analysis. The counts are simply the absolute amount of each read while the abundances are a normalized quantification of the reads, measured in transcripts per million (TPM).

[5] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

[6] Denoeud F, Carretero-Paulet L, Dereeper A et al, " The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.", 2014, Science, 345(6201):1181-1184.

[7] Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology **34**, 525–527, 2016, doi:10.1038/nbt.3519R Core

Statistical Evaluation and Differential Expression Analysis

The abundance files generated by kallisto were imported into R[8] using the package tximport[9] and formatted using tidyverse[10]. Ensembl annotations were imported using biomaRt[11] and ensembldb[12].

The counts were then transformed into a DGEList object using edgeR[13] to map transcript IDs to gene IDs, which was then filtered to only include genes which had been expressed in at least one sample. Additionally genes that had a total count below 1 CPM over any 4 samples were removed. Next the DGEList object was normalized to counts per million (CPM), log2 transformed and normalized using the trimmed mean of M-values (TMM) method to adjust for differences in library size and gene length across samples.

The packages matrixStats[14] and cowplot[15] were additionally used to visualize the outputs (see supplemental material: https://github.com/vm-vh/RNASeq_Analysis/blob/main/1_import_filter_norm.R).

To find possible patterns in the filtered and normalized data hierarchical clustering and principal component analysis (PCA) were performed. Hierarchical clustering was done using both the "euclidean" and "maximum" distance methods for comparison. PCA was performed using the "prcomp" function, and the resulting eigenvalues were used to calculate the percentage of variance explained by each principal component.

Additionally average expression levels for each gene were calculated in both stem and leaf tissues and plotted against each other.

The next step was to identify differentially expressed genes (DEGs) using linear modeling and hypothesis testing. The samples were grouped based on tissue type and the 'limma' package was then used to model the mean-variance trend and fit a linear model to the data. The means of two sample groups were compared and Bayesian statistics were calculated for the linear model fit.

The all differentially expressed genes (DEGs) were classified as upregulated, downregulated, or not significantly different between the two tissues, based on a specified significance level (adjusted p-value < 0.05) and fold change (log2 fold change > 1 or < -1). A

[8] https://www.R-project.org/.

[9] Charlotte Soneson, Michael I. Love, Mark D. Robinson, “Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences”, 2015, F1000Research

[10] Wickham H et al, "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 2019, 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

[11] Robinson MD, McCarthy DJ and Smyth GK,. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.”, 2010, Bioinformatics 26, 139-140

[12] Rainer J, Gatto L, Weichenberger CX, “ensembldb: an R package to create and use Ensembl-based annotation resources”., 2019, Bioinformatics., doi:10.1093/bioinformatics/btz031

[13] Robinson MD, McCarthy DJ and Smyth GK, “ edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”, 2010,. Bioinformatics 26, 139-140

[14] Bengtsson H (2022). _matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)_. R package version 0.63.0, <https://CRAN.R-project.org/package=matrixStats>.

[15] Wilke C (2020). _cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'_. R package version 1.1.1, <https://CRAN.R-project.org/package=cowplot>.

volcano plot was then generated to visualize the DEGs, with genes plotted based on their log2 fold change and negative log10 adjusted p-value.

Lastly, the rows and columns of the gene expression data were clustered using the Pearson correlation coefficient for the rows and the Spearman correlation coefficient for the columns. Each gene was then assigned to one tissue module using the clusters obtained from row clustering and heatmaps were generated to visualize the gene expression data overall as well as in the two tissue modules with the clustered genes as rows and the samples as columns.

Gene Ontology (GO) enrichment analysis was performed on the samples using the gProfiler2[16] package. First, the top 100 genes based on their log-fold change values were selected to perform GO enrichment analysis. GO enrichment analysis was additionally performed on the genes belonging to each tissue module separately.

The False Discovery Rate correction method ("fdr") for multiple testing was used for all three enrichment analyses and the results were plotted using interactive Manhattan plots (see supplemental material: https://github.com/vm-vh/RNASeq_Analysis/blob/main/2_analysis.R).

Results

Quality Evaluation

The FastQC reports of the original data each had the same two warnings and one failure. The reports generated warnings in the categories “Per Base Sequence Content”, which analyzes the percentage of each base at each position in the sequence and issues a warning if the difference between A and T, or G and C is greater than 10% at any position, and “Sequence Length Distribution”, which raises a warning when all sequences are not the same length. The category “Sequence Duplication levels” failed which means that more than 50% of sequences were not unique.

After trimming the data using Trimmomatic the same failure and warnings persisted in the report. However since the overall quality of the data was relatively good from the beginning it was decided to continue without trimming again as the loss of information would have been too great in comparison to the achieved quality enhancement.

Mapping Efficiency and Coverage

The percentage of fragments that could be pseudoaligned (p_pseudoaligned) was each ca. 80% for the leaf samples and ca. 85% for the stem samples. The percentage of fragments

[16] Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H,. "gprofiler2- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler." , 2020, , _F1000Research_, *9 (ELIXIR)*(709). R package version 0.2.1.

that kallisto was able to uniquely align to a target sequence (p_unique) were only slightly lower at ca 79% for the leaf samples and ca. 84% for the stem samples.

Normalization and filtering

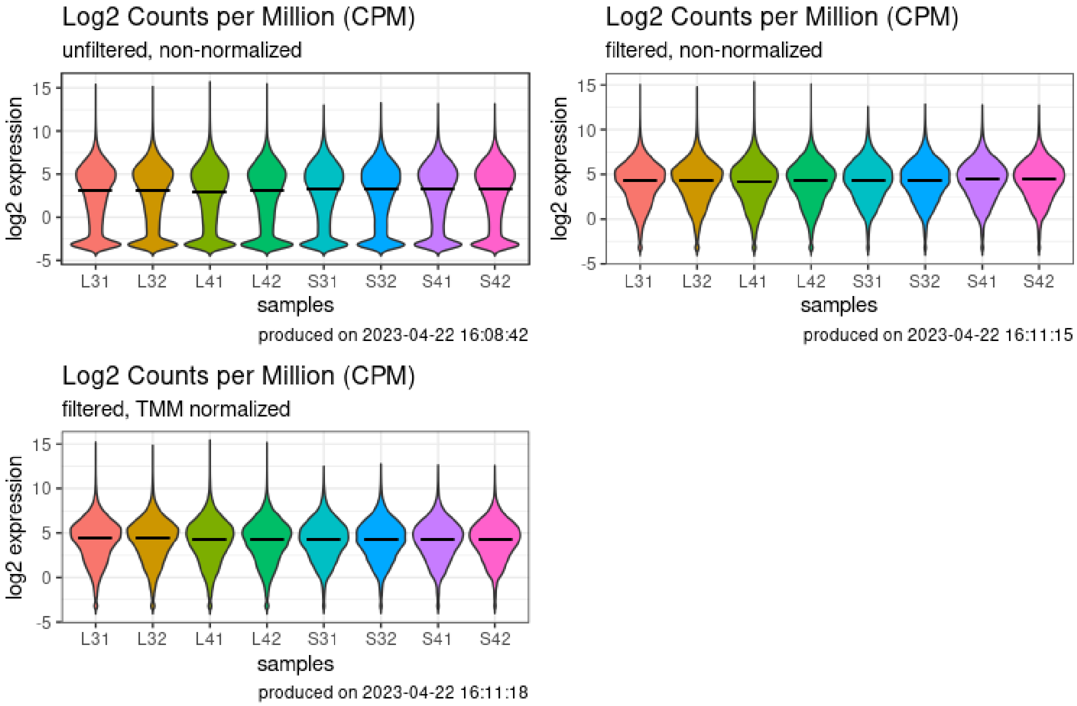


Figure 1. Violin plots of expression values in log2 CPM for each sample in an unfiltered, non-normalized state, a filtered, non-normalized state and a filtered, normalized state.

As can be seen in Fig. 1, filtering of the gene expression data was successful in removing genes with very low expression values. Normalization on the other hand seems to not have had much of an effect, it is therefore likely that the data had already been normalized.

Exploratory Data Analysis

The *C. canephora* plants were grown naturally in the Dehong Tropical Agriculture Research Institute in Yunnan Province, China. Therefore, local environmental conditions could have influenced the plants, but likely would have influenced all plants equally. Several different plants and replicates of each sample were used to guard against outliers.

The clusters generated by the hierarchical clustering matched the origin of the samples, both methods could clearly distinguish between leaf and stem tissues. Within the individual tissues both methods were able to correctly separate the leaf samples originating from separate plants however only the euclidean method was able to correctly separate the stem samples (Fig. 2). This suggests that the stem samples might be more similar to each other compared to the leaf samples.

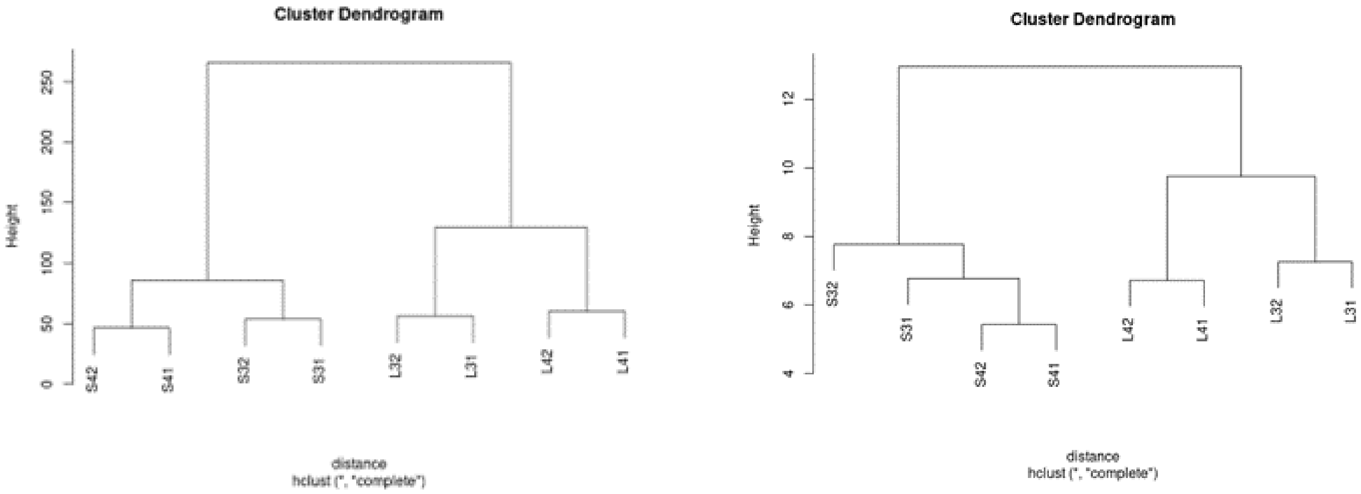


Figure 2: Two cluster dendrograms, the left dendrogram was generated using the euclidean distance method while the right dendrogram was generated using the maximum distance method .

The results of the principal component analysis (Fig. 3 and 4) match well with the results of the dendrograms in that there is a clear distinction between leaf and stem samples. The PCA also clustered the stem samples closer together compared to the leaf samples.

The two plotted principal components (PCs) explain a combined 92.3% of the variance between the samples. PC1 describes 82.5% of the variance and distinguishes mostly between leaf and stem tissue while PC2 describes 9.8% of the variance and shows the differences within the samples of each tissue type.

According to the small multiples PCA plot (Fig. 4) the third PC explains most of the variance between the stem samples and the fourth PC captures the variance between the individually sampled plants.

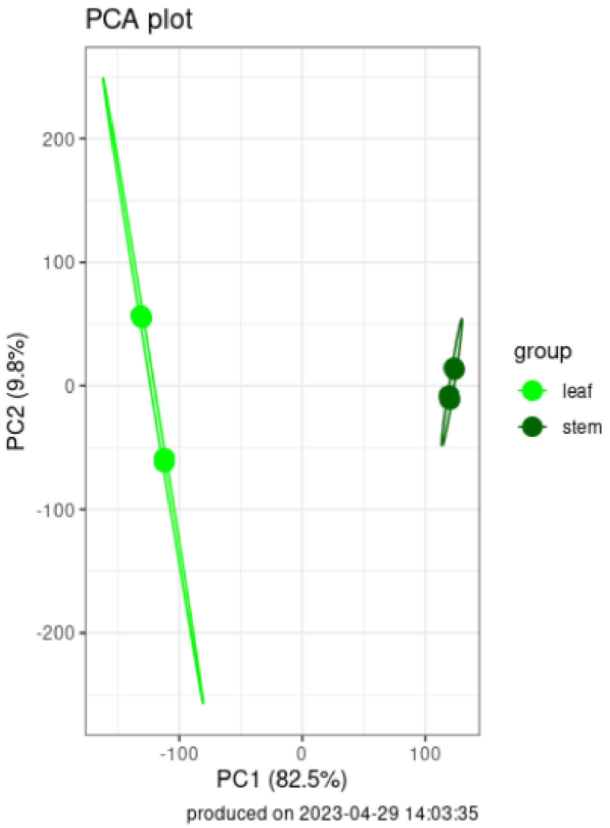


Figure 3: A PCA plot with the groups leaf and stem

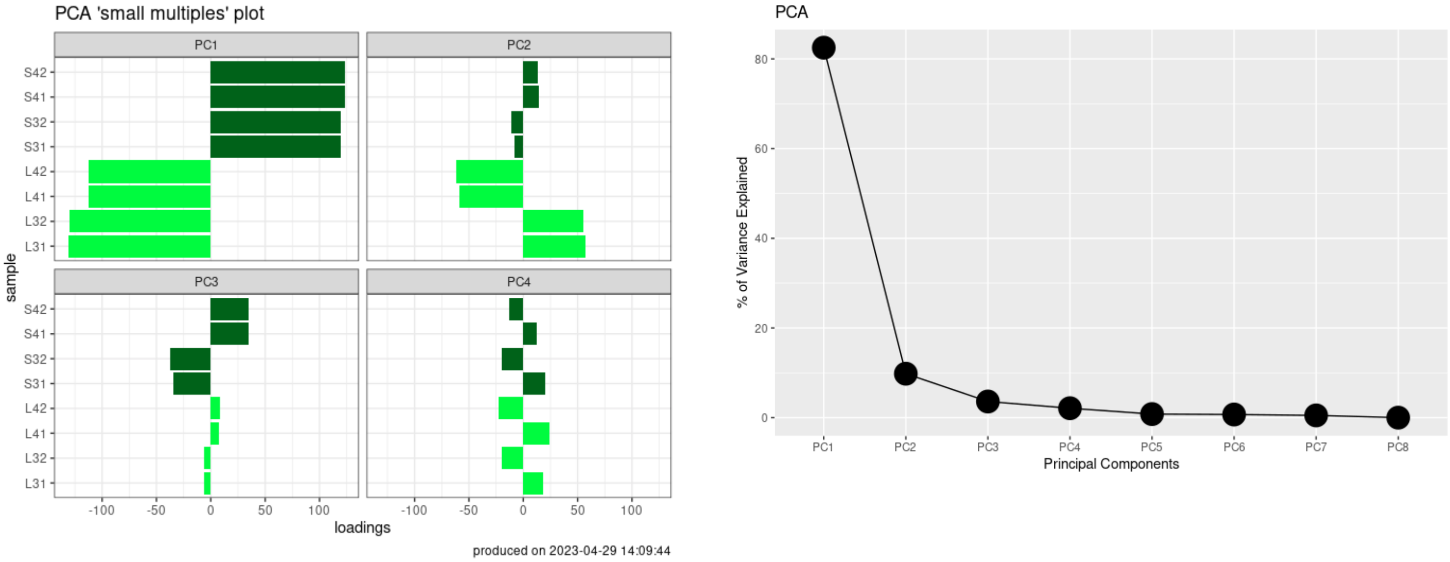


Figure 4: A small multiples PCA plot (left) and a plot of all 8 principal components (right).

The resulting graph of plotting each gene's average expression level in leaf and stem tissue (Fig. 5) shows a rough correlation.

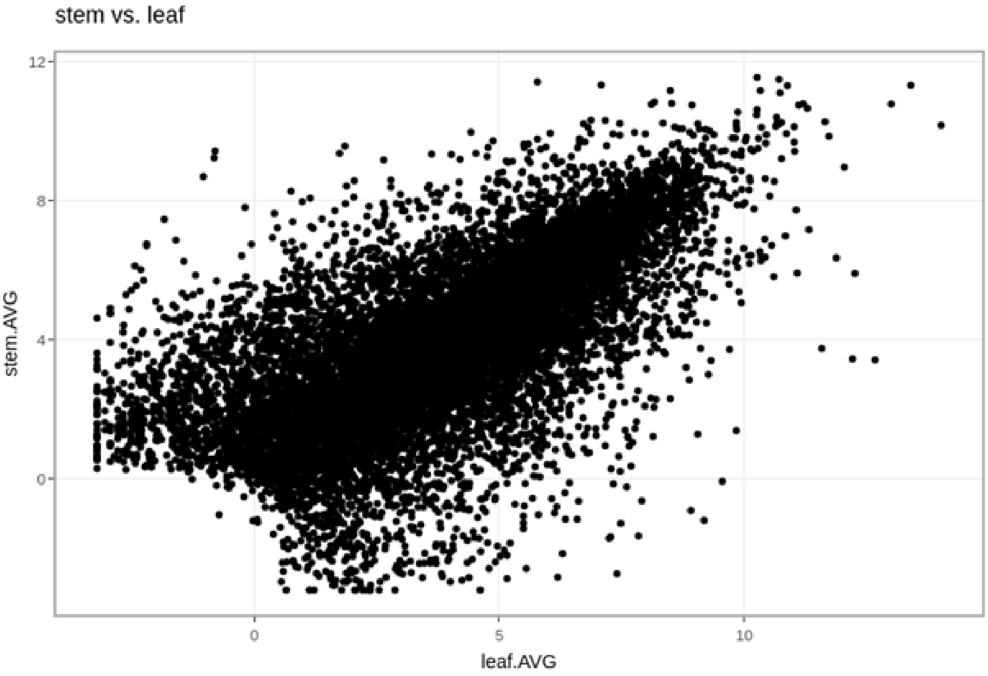


Figure 5: A scatterplot of the average expression levels of all genes in the leaf tissue plotted against the average expression levels of all genes in the stem tissue.

This result makes sense because while this study will focus on differentially expressed genes, a majority of the expressed genes in these tissues should be housekeeping genes which are needed in almost all cells to maintain fundamental functions. For differential gene expression analysis it is promising that there seem to also be a fair number of genes which show very different average expression values between the two tissues.

Differentially Expressed Genes

A total of 6.885 significantly differentially expressed genes could be identified. Nine genes were chosen for further interpretation (Table 1) based on either their low adjusted p-value or an extreme fold change (FC) value. Five of the selected genes were chosen from the dark green region in the volcano plot (Fig. 6) which corresponds to a low FC and therefore to the stem tissue while the other four genes were chosen from the light green region in the volcano plot (Fig. 6) which corresponds to a high FC and therefore to the leaf tissue.

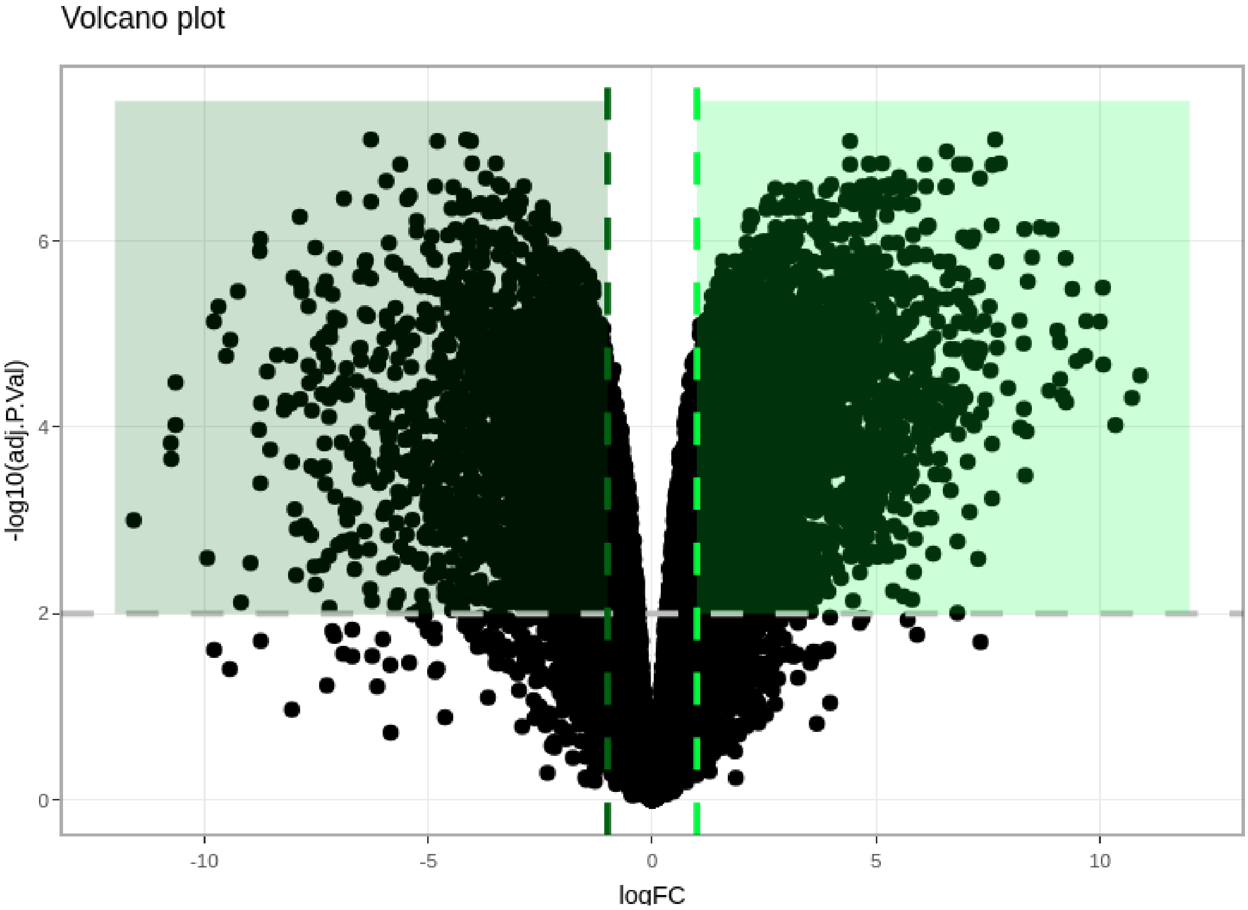


Figure 6: A volcano plot of all genes, significantly differentially expressed genes are highlighted in dark green (stem tissue) and light green (leaf tissue).

Gene ID	logFC	-log10(adj.P.Val)	expression /tissue	GO annotations
GSCOC_T00022211001	-1.06e+01	4.47e+00	under/stem	CC: membrane[17]
GSCOC_T00037367001	-1.06e+01	4.34e+00	understem	CC: extracellular region[18]
GSCOC_T00028630001	-6.24e+00	7.09e+00	under /stem	CC: apoplast MF: hydrolase activity MF: xyloglucan BP: cell wall organization BP: cellular glucan metabolic process[19]
GSCOC_T00012375001	-4.76e+00	7.09e+00	under /stem	CC: membrane MF: metal ion transmembrane transporter activity[20]
GSCOC_T00029585001	-4.01e+00	7.09e+00	under /stem	CC: membrane MF: ABC-type transporter activity M: ATP binding[21]
GSCOC_T00018057001	4.46e+00	7.09e+00	over /leaf	CC: membrane MF: oxidoreductase activity BP: lipid metabolic process[22]
GSCOC_T00031499001	6.61e+00	7.09e+00	over /leaf	CC: apoplast MF: manganese ion binding[23]

[17] https://www.uniprot.org/uniprotkb/A0A068TPI8/entry
[18] https://www.uniprot.org/uniprotkb/A0A068UWT4/entr
[19] https://www.uniprot.org/uniprotkb/A0A068ULJ6/entry
[20] https://www.uniprot.org/uniprotkb/A0A068VHE3/entry
[21] https://www.uniprot.org/uniprotkb/A0A068TV10/entry
[22] https://www.uniprot.org/uniprotkb/A0A068V6X3/entry
[23] https://www.uniprot.org/uniprotkb/A0A068UQM7/entry

GSCOC_T00021809001	7.69e+00	7.09e+00	over /leaf	MF: fatty acid binding BP: systematic acquired resistance[24]
GSCOC_T00012163001	1.09e+01	4.59e+00	over /leaf	CC: apoplast MF: copper ion binding MF: hydroquinone BP: lignin catabolic process[25]

Table 2. Genes with a low adjusted p-value or an extreme FC value.

The biological significance of these genes is captured in their gene ontology (GO) annotations. Many over-/under-expressed genes have responsibilities within membrane related processes.

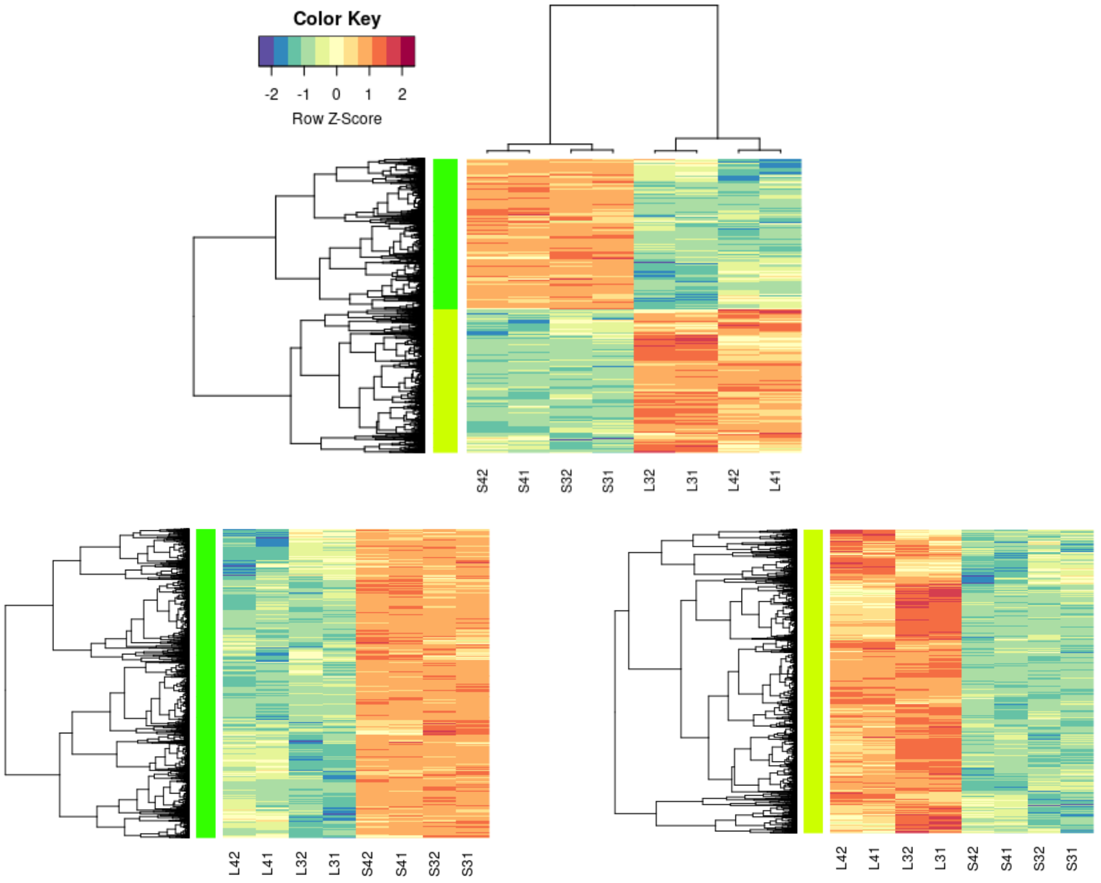


Figure 7: A heatmap showing differentially expressed genes with the colors on the left side corresponding to the stem (dark green) and leaf (green) tissue modules (top), a heatmap showing differentially expressed genes within

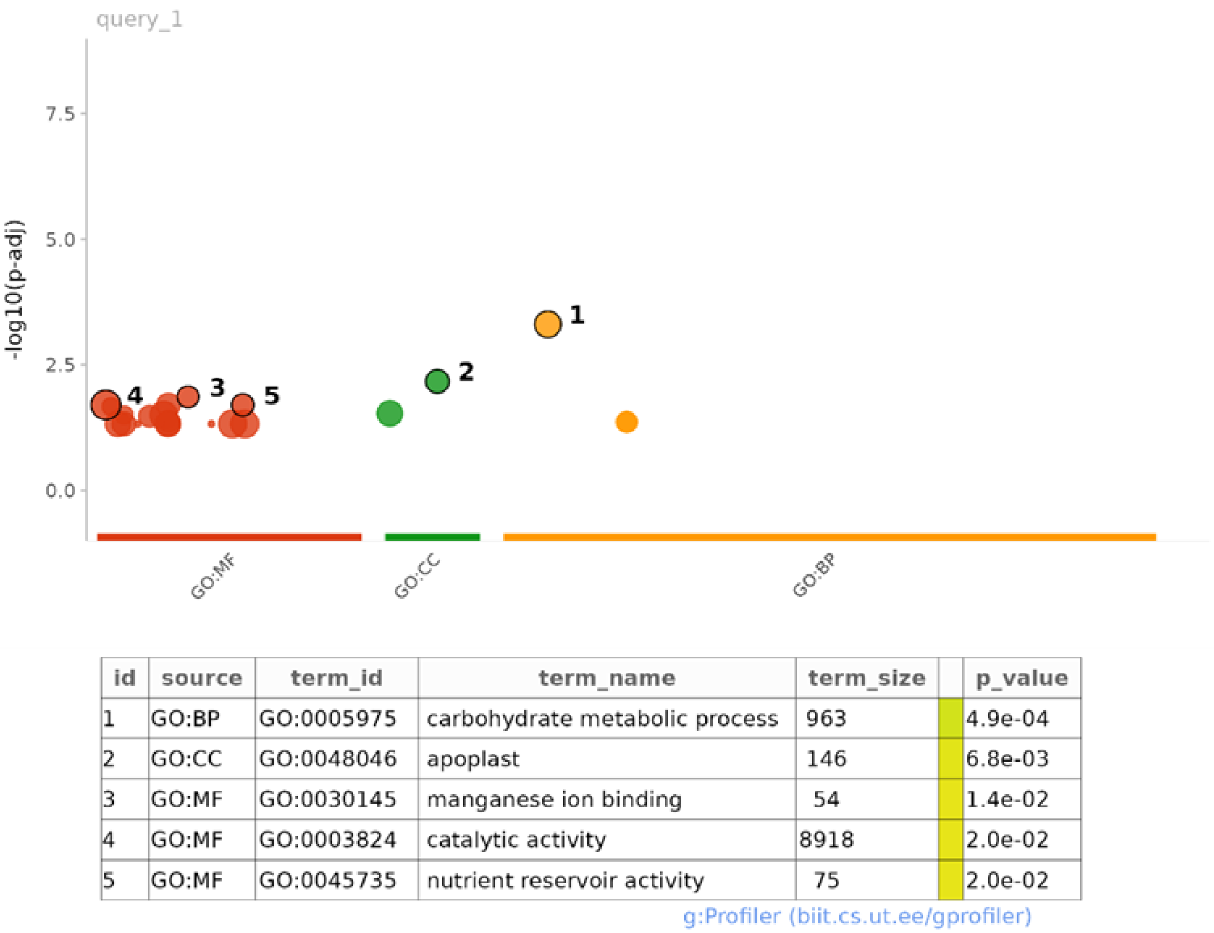
[24] https://www.uniprot.org/uniprotkb/A0A068TR77/entry

[25] https://www.uniprot.org/uniprotkb/A0A068VJQ7/entry

the leaf tissue module (bottom left) and a heatmap showing differentially expressed genes within the stem tissue module (bottom right)

To visualize all differentially expressed genes, a heatmap was generated (Fig. 7). Two additional heatmaps (Fig. 7) were generated for the leaf and stem tissue modules separately. Within the leaf tissue module, differentially expressed genes within the leaf samples generally have Z-scores above 0 while the DEGs in the stem samples generally have Z-scores below 0. For the stem tissue module the exact opposite pattern can be observed with DEGs in leaf samples generally having Z-scores below 0 while the DEGs in the stem samples generally have Z-scores above 0.

Functional Enrichment Analysis

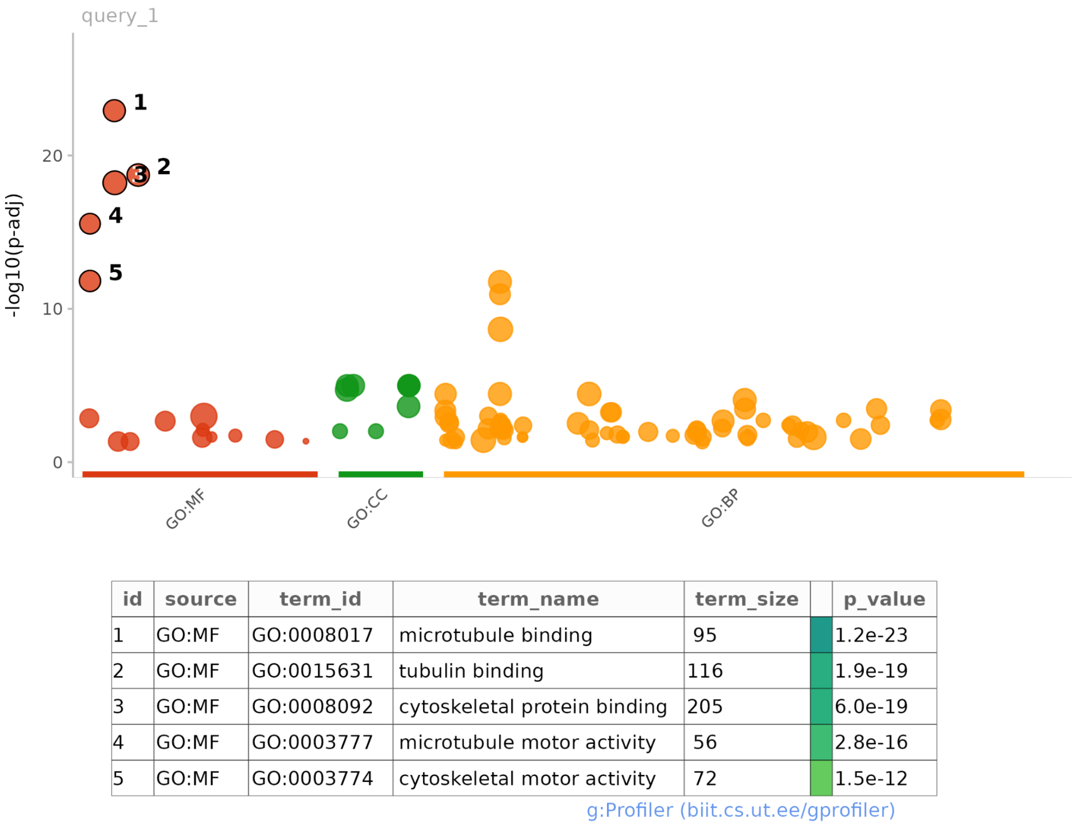


id	source	term_id	term_name	term_size		p_value
1	GO:BP	GO:0005975	carbohydrate metabolic process	963		4.9e-04
2	GO:CC	GO:0048046	apoplast	146		6.8e-03
3	GO:MF	GO:0030145	manganese ion binding	54		1.4e-02
4	GO:MF	GO:0003824	catalytic activity	8918		2.0e-02
5	GO:MF	GO:0045735	nutrient reservoir activity	75		2.0e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 8: GO enrichment plot and first 5 rows of the corresponding table (sorted by p-value) for the top 100 most differentially expressed genes.

Performing functional enrichment analysis for the top 100 differentially expressed genes based on FC only yields relatively broad results (Fig. 8), catalytic activity and carbohydrate metabolic processes especially are very non-specific annotations. The most significantly enriched annotations are not limited to one GO category (biological processes, cellular components and molecular functions). This is likely because the analysis is done for both types of tissues at once. This suggested that further analysis is necessary to understand the differences in GO annotations between the two tissues of interest.



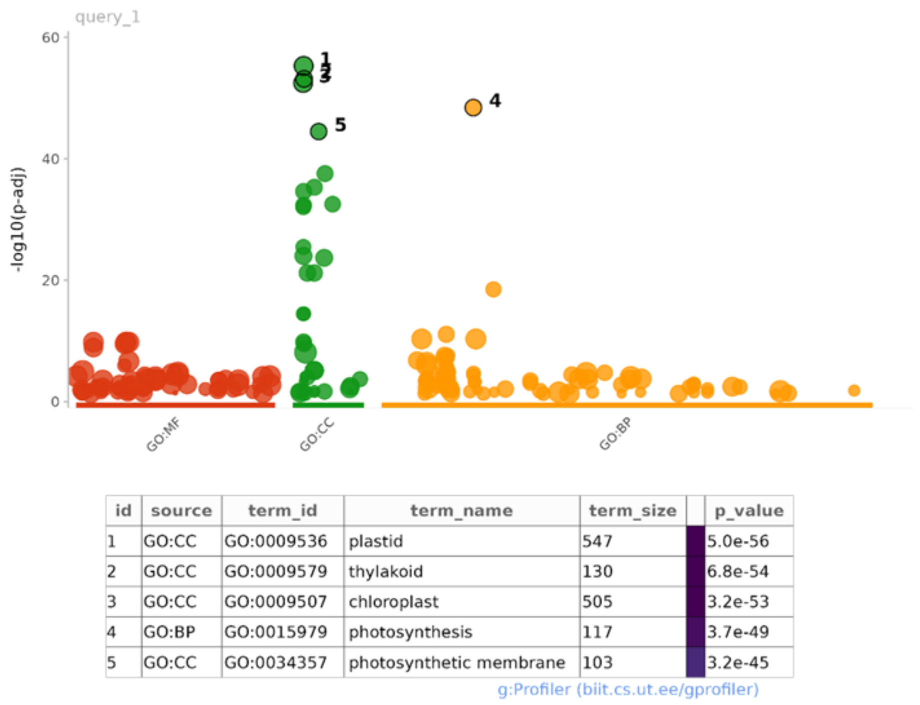
id	source	term_id	term_name	term_size		p_value
1	GO:MF	GO:0008017	microtubule binding	95	🟩	1.2e-23
2	GO:MF	GO:0015631	tubulin binding	116	🟩	1.9e-19
3	GO:MF	GO:0008092	cytoskeletal protein binding	205	🟩	6.0e-19
4	GO:MF	GO:0003777	microtubule motor activity	56	🟩	2.8e-16
5	GO:MF	GO:0003774	cytoskeletal motor activity	72	🟩	1.5e-12

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 9: GO enrichment plot and first 5 rows of the corresponding table (sorted by p-value) for all significantly differentially expressed genes within the stem tissue.

GO enrichment for all significantly differentially expressed genes within the stem tissue yielded more expected results since all of the five most significantly enriched annotations are related to the cytoskeleton within the GO category ‘molecular functions’. These enriched annotations align with the biological role of the stem in plant stability and transport.

The GO enrichment for all significantly differentially expressed genes within the leaf tissue also matched the expectations as they did in the stem tissue. The five most significantly enriched annotations are related to photosynthesis and are therefore within the GO categories ‘cellular components’ and ‘biological processes’. These significantly enriched GO annotations align with the biological role of the leaf as the primary site of photosynthesis in plants.



id	source	term_id	term_name	term_size		p_value
1	GO:CC	GO:0009536	plastid	547	▇	5.0e-56
2	GO:CC	GO:0009579	thylakoid	130	▇	6.8e-54
3	GO:CC	GO:0009507	chloroplast	505	▇	3.2e-53
4	GO:BP	GO:0015979	photosynthesis	117	▇	3.7e-49
5	GO:CC	GO:0034357	photosynthetic membrane	103	▇	3.2e-45

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 10: GO enrichment plot and first 5 rows of the corresponding table (sorted by p-value) for all significantly differentially expressed genes within the leaf tissue.

Discussion

Overall the results indicated that the data was of good quality and reliability.

The significantly enriched GO annotations in the stem tissue related to the cytoskeleton, which aligns with the biological role of the stem in plant stability and transport. Similarly, the significantly enriched GO annotations of the leaf tissue related to photosynthesis, which aligns with the biological role of the leaf as the primary site of photosynthesis in plants. These results confirm the existing knowledge in the scientific community, support the work of many other researchers and again confirm that the data was reliable and of good quality.

While this study succeeded in its primary goal of performing a successful RNASeq analysis most of the obtained results are not new or surprising as differences in tissue types in different plant species have been extensively studied. This was very good for the primary goal of this study as it is therefore possible to say with a high amount of certainty that the RNASeq analysis worked correctly and that the data was of good quality. However, for results that could contribute something to the scientific community, several parameters would need to be changed, making much more time and equipment necessary.

One issue is that this analysis included too few samples and replicates, so to enhance the statistical significance of our results, more samples and replicates should be analyzed. Another problem is that the growth of the plants and the RNA extraction were not controlled by our scientific team but by an external team. Therefore the detailed environmental growth conditions could not be controlled sufficiently. These issues would need to be addressed to obtain more meaningful results that could contribute to the scientific community.

The original study aimed to compare caffeine production in coffee and tea plants, which is an interesting idea but would require a more complicated RNASeq analysis as well as further analysis.

Conclusion

The RNA-seq data analysis presented in this study has provided insight into the genetic basis of the differences between the stem and leaf tissues of *C. canephora*. This study identified significant variations in the transcriptomes of the two tissues, with over-/under-expressed genes indicating that the leaf tissue primarily functions in sugar production through photosynthesis, while the stem tissue is responsible for stability and transport-related processes.

The findings of this study confirm the existing knowledge in the scientific community about the biological roles of these plant tissues. Plants rely on their leaves for photosynthesis, which involves capturing light energy and converting it into chemical energy in the form of sugars. In contrast, the stem provides support and transports water, nutrients, and other essential substances between the roots and the leaves. The results of this study, therefore, align with the known biological functions of the two tissues.

Supplemental Information

All R scripts, error logs and results at https://github.com/vm-vh/RNASeq_Analysis

References

Noirot M, Poncet V, Barre P, Hamon P, Hamon S, De Kochko A. 2003. Genome size variations in diploid African Coffea species. Annals of Botany 92: 709-714.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo: g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) Nucleic Acids Research 2019; doi:10.1093/nar/gkz369 [PDF].

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H(2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686

Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research

Rainer J, Gatto L, Weichenberger CX (2019) ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics. doi:10.1093/bioinformatics/btz031

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).

Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140

Bengtsson H (2022). _matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)_. R package version 0.63.0, https://CRAN.R-project.org/package=matrixStats

Wilke C (2020). _cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'_. R package version 1.1.1, https://CRAN.R-project.org/package=cowplot

Xie Y, Cheng J, Tan X (2023). _DT: A Wrapper of the JavaScript Library 'DataTables'_. R package version 0.27, https://CRAN.R-project.org/package=DT

C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2023). _gt: Easily

Create Presentation-Ready Display Tables_. R package version 0.9.0, https://CRAN.R-project.org/package=gt

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Tal Galili, Alan O'Callaghan, Jonathan Sidi, Carson Sievert; heatmaply: an R package for creating interactive cluster heatmaps for online publishing, Bioinformatics, , btx657, https://doi.org/10.1093/bioinformatics/btx657

Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2022). _gplots: Various R Programming Tools for Plotting Data_. R package version 3.1.3, https://CRAN.R-project.org/package=gplots

Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H (2020). "gprofiler2- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler." _F1000Research_, *9 (ELIXIR)*(709). R package version 0.2.1.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology **34**, 525–527 (2016), doi:10.1038/nbt.3519R Core

Team R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
(2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Zhang, Yanrui & Fu, Jiamin & Zhou, Qiying & Li, Fangdong & Shen, Yihua & Ye, Zhili & Tang, Dingkun & Chi, Ning & Li, Lanqing & Ma, Shuyu & Inayat, Ali & Guo, Tieying & Zhao, Jian & li, Penghui. (2022). Metabolite Profiling and Transcriptome Analysis Revealed the Conserved Transcriptional Regulation Mechanism of Caffeine Biosynthesis in Tea and Coffee Plants. Journal of Agricultural and Food Chemistry. 70. 10.1021/acs.jafc.1c06886.

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA798825

The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.
Denoeud F, Carretero-Paulet L, Dereeper A et al. . 2014. Science. 345(6201):1181-1184.

https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

https://www.uniprot.org/uniprotkb/A0A068TPI8/entry

https://www.uniprot.org/uniprotkb/A0A068UWT4/entry

https://www.uniprot.org/uniprotkb/A0A068ULJ6/entry

https://www.uniprot.org/uniprotkb/A0A068VHE3/entry

https://www.uniprot.org/uniprotkb/A0A068TV10/entry

https://www.uniprot.org/uniprotkb/A0A068V6X3/entry

https://www.uniprot.org/uniprotkb/A0A068UQM7/entry

https://www.uniprot.org/uniprotkb/A0A068TR77/entry

https://www.uniprot.org/uniprotkb/A0A068VJQ7/entry