Düsseldorf, 01.05.2023

The comparison of RNA sequences between flower and fruit tissue of Solanum Lycopersicum.

Franka Prinz, Natalia Rekiel

Abstract

For this assignment we made a comparison between replicates of flower tissue and fruit tissue of Solanum Lycopersicum, to check whether there are any significant differences. For this purpose we used FastQC, Ubuntu and RStudio. FastQC helped us check the quality of the chosen sequences and monitor it while preprocessing; Ubuntu was used for preprocessing of our sequences (Trimmomatic), sequencing reads to a reference genome (HISAT2), converting our .sam files to .bam files (samtools), and pseudomapping (Kallisto). The results of pseudomapping were further analysed in RStudio, which made us prove that there is a significant difference between a flower and fruit tissue of Solanum Lycopersicoides.

Introduction

Solanum Lycopersicum, well known as tomato, provides nutrient rich fruits which are present in our every day life. It is also a very well-studied organism, providing a lot of data for further analysis. Because of that we decided to perform a comparison of two different tissues of this plant. The chosen tissues are: flower and fruit tissue. The chosen method is RNA sequencing, as this method is unbiased, has much higher coverage, and provides very accurate measurements of gene expression, even for genes expressed at very high or low levels while requiring low RNA amounts. This method provides insight to the transcript number of genes, what genes are transcribed at specific time, helps identify biomarkers, as well as it informs the researchers about alternative splicing, transcript variation or novel transcripts. This method is attractive to study plants, as it can give an insight to which certain genes are activated or what their expression is. It provides a better understanding of plant life in general, the plant stress, and with further investigation can help improve crop yield. Our part of this study is to evaluate the quality (FastQC), map the sequences to a respective genome (HISAT2&Kallisto), perform statistical evaluation (RStudio), differential expression analysis(RStudio), functional enrichment analysis (RStudio), and evaluate and discuss our results.

Materials and methods

Selection of Publicly Available Plant RNA-seq Data

For this study, three RNA-seq replicates of each flower and fruit tissue of Solanum Lycopersicoides were chosen. They are coming from the same project, and same biosample to ensure that there are no differences in the methods used and that the library was run in single-end mode.

The other requirement was that they should have a similar size. The replicates were downloaded from National Center for Biotechnology Information (NCBI), project: PRJNA727176, replicates: SRR17655903, SRR17655904, SRR17655908 (flower) and SRR17655900, SRR17655901, SRR17655902 (fruit).

Quality Evaluation

Firstly, we performed a quality control to ensure that we use high quality data for our study and that our results will be significant. To check the quality of chosen replicates, we ran a quality control using FastQC, which is “a quality control tool for high throughput sequence data”.

Mapping to the Respective Genome

We performed mapping and pseudomapping of our replicates to the reference genome:

⇒ Mapping - HISAT2

For mapping of our replicates we used HISAT2, which is an alignment for mapping next-generation sequencing reads. The reference genome used for this purpose was downloaded from solgenomics.net, the version of the tomato genome was SL2.0.

⇒ Pseudomapping - Kallisto

For pseudomapping of our replicates we used Kallisto. We used a reference genome from plants.ensembl.org. As length and standard deviation we put 180 and 20, respectively. The results were further investigated in RStudio.

Statistical Evaluation and Differential Expression Analysis

We used R-Studio to evaluate the data and to identify the differentially expressed genes. To do that we created a text file containing the names of the files we got as a result of the pseudomapping, the sample names, and the respective group.
We used this file to create a path that leads us to the quantification files, which is what we’re interested in. The quantification files (abundance.tsv) contained the data on the gene expression in the respective sample. Using the Solanum Lycopersicum genome from bioMart we quantified the abundance of each gene in each sample.

```
Txi_gene <- tximport(path,
                     type = "kallisto",
                     tx2gene = Tx.lyc,
                     txOut = FALSE, #How does the result change if this =FALSE vs =TRUE?
                     countsFromAbundance = "lengthScaledTPM",
                     ignoreTxVersion = FALSE)

Txi_gene
```

We created a first plot of our acquired data by transforming the data using the function transform. We created a DGEList (Digital Gene Expression List) containing the counts matrix with the raw read counts for each gene in each sample, the gene IDs and the sample names as well as the total number of reads for each sample (library size). We normalized the DGEList using CPM (counts per million) which is scaling the counts by library size to ensure a comparison between two samples with a different library size is possible. We had to transform our data from wide format to long format using pivot, so we would only have three columns containing all the information. This way we were able to compare different groups over time using a plotting function. Next we filtered our data and got rid of the data points with low transcript representation and normalized our data again using CPM. We got the normalization factors for our data using

“calcNormFactors()” and stored them in a list, which we used later.

```
myDGEList.filtered.norm <- calcNormFactors(myDGEList.filtered, method = "TMM")
log2.cpm.filtered.norm <- cpm(myDGEList.filtered.norm, log=TRUE)
log2.cpm.filtered.norm.df <- as_tibble(log2.cpm.filtered.norm, rownames = "geneID")
colnames(log2.cpm.filtered.norm.df) <- c("geneID", sampleLabels)

log2.cpm.filtered.norm.df.pivot <- pivot_longer(log2.cpm.filtered.norm.df,
                                                cols = FL01:FR03,
                                                names_to = "samples",
                                                values_to = "expression")
```

From this normalized and filtered data we created a dendrogram cluster, showing the similarities and differences between our samples. The length of the line connecting two samples shows how similar these are.

The next step was to find out the principal components affecting our samples. For this we used PCA (principal component analysis). PCA reduces the dimensionality of our data while keeping as much of the data points as possible. We only used the first two principal components for our plots, since those have the most notable effect on our data.

Instead of looking at the values of each sample separately we created a datatable containing the average values for each group. This gives us information on the average expression of genes in our two groups (flower and fruit).

We designed a matrix to take a look at the relation between flower and fruit using a volcano plot. A volcano plot represents the significance of the gene (p-value) and magnitude of the difference in expression levels. The points far up on the left or right represent genes that are significant and show a distinct difference in expression between the groups we picked. We filtered the results and kept only the ones with a p-value higher than 0.05. At the end of this we have a list of 50 genes that are differentially expressed and have a significant p-value.

Functional Enrichment Analysis

```
myTopHits <- topTable(ebFit, adjust ="BH", coef=1, number=50, sort.by="logFC")
nrow(myTopHits)
myTopHits
#also used website
gost.res_up <- gost(rownames(myTopHits), organism = "slycopersicum", correction_method = "fdr")
```

We used the list of genes (MyTopHits) to perform a gene ontology enrichment analysis using gprofiler. The myTopHits list, containing our genes of interest, was based on the DGEList created in one of the first steps. We used the voom function on the list to estimate the weight of the observation. The voom function does this by looking at the mean-variance relation. We did it once in R and once on the gprofiler website.

Results

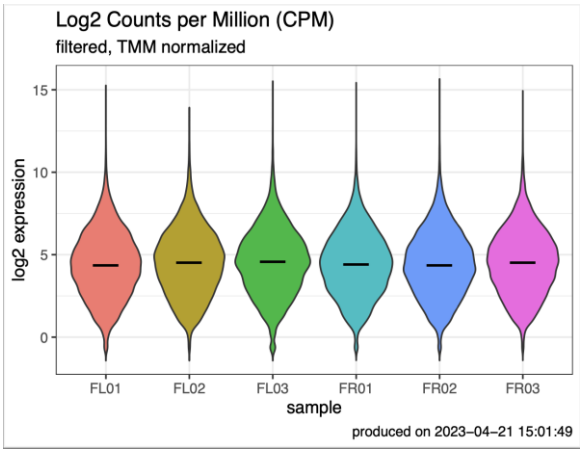
Quality Evaluation

The biggest problem during the quality control was the GC content. It turned out that the replicates had a Poly-A tail at the beginning. We trimmed our sequences by cropping 60 bases off the front of each replicate. Besides that, we removed low-quality bases from single-end RNA-seq reads, and made sure to keep reads of length at least 70. This significantly improved the quality of our replicates.

Mapping Efficiency and Coverage

The results vary from 76.85% to 90.42%, which is a very good mapping rate, ensuring us that the quality evaluation we performed correctly, the reference genome is high quality and that there is no contamination.

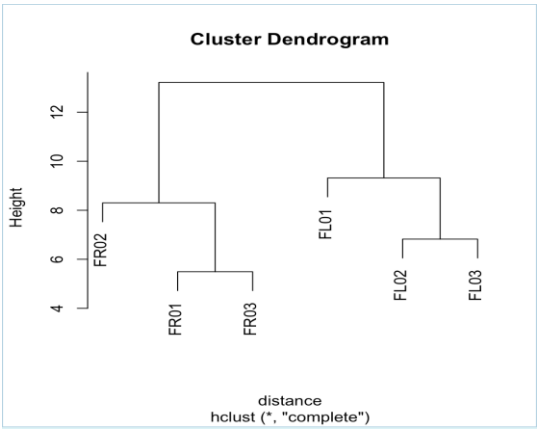
Exploratory Data Analysis



Plot 1 - CPM

Our normalized and filtered data shows (Plot 1 - CPM) that we have approximately a normal distribution within our samples and the overall expression levels seem to be similar. This is an important factor for the further analysis of our data.

The cluster dendrogram shows the differences in expression between our samples. As you can see (Plot 2 – Cluster Dendogram) the two groups flower (FL01, FL02, FL03) and fruit (FR01, FR02, FR03) are clustered together meaning that the differences in gene and transcript expression within these groups are lower than the differences between two samples from different groups. This aligns with our assumption that samples from the same tissue would have similar gene expression levels. FR01 and FR03 seem to be more closely related than FR02 to one of them. The reason for that is unclear. It could be that FR01 and FR03 are in the exact same state of development, while FR02 is in a different state. But it is difficult to make assumptions about that.

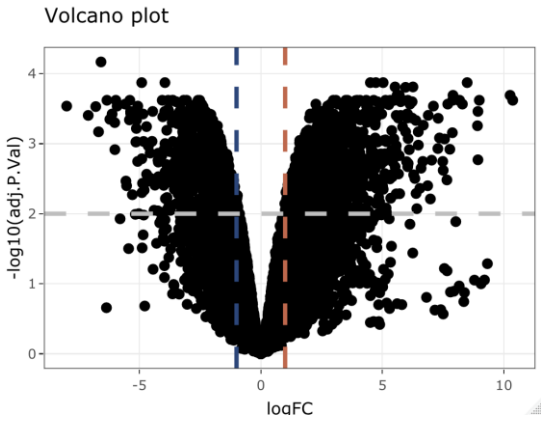


Plot 2 – Cluster Dendogram

The PCA (Plot 3 – PCA) showed that there were four principal components. The first two had a mentionable effect on our data and were thus included in the further analysis.

Plot 3 – PCA

Differentially Expressed Genes



Plot 4 – Volcano Plot

As you can see in the volcano plot (Plot 4 – Volcano Plot) our two groups (flower and fruit) have many significantly differentially expressed genes. This aligns with our assumption that the gene expression in flower and fruit is different. We only took the top fifty genes, with the highest p-value. If we take a closer look at our MyTopHits list we find that the top gene is Solyc12g005940.2. It is a gene that plays part in a metabolic pathway.

```
> myTopHits
                  logFC  AveExpr         t      P.Value    adj.P.Val          B
Solyc12g005940.2 10.364248 4.639533  14.006110 4.227279e-07 0.0002399857  5.66624652
Solyc08g067230.3 10.255371 6.430505  15.684809 1.691491e-07 0.0002039120  7.04557891
Solyc01g066380.3  9.322709 6.711469   2.965013 1.712942e-02 0.0516752524 -3.34087863
Solyc01g056310.3  9.201225 5.664491   2.496650 3.590036e-02 0.0883794871 -3.89552580
Solyc01g068120.3  9.058864 5.266454   2.388984 4.260700e-02 0.1001206197 -4.01077763
Solyc03g113560.3  8.987627 5.110700  13.328674 6.299007e-07 0.0002399857  5.81347951
Solyc09g014550.3  8.932511 4.936262   7.158065 7.689287e-05 0.0016984300  2.03865353
Solyc08g067030.3  8.932194 6.193820  11.376513 2.226360e-06 0.0003462903  5.31751652
Solyc02g093180.3  8.912823 5.876051   9.571854 8.604709e-06 0.0005547595  4.09045100
Solyc12g005320.2  8.788543 5.556246   2.493122 3.610230e-02 0.0887026829 -3.92224193
Solyc01g099630.3  8.489167 6.047350  17.769286 6.119135e-08 0.0001343231  8.01368385
Solyc03g058910.3  8.377322 4.808574   2.142983 6.298241e-02 0.1334798433 -4.35298321
Solyc06g084620.1  8.350363 4.273456   1.884654 9.458688e-02 0.1798893026 -4.64720647
Solyc12g006050.2  8.292159 5.406878  12.471310 1.072665e-06 0.0002722326  5.74092215
Solyc10g075110.2  8.177093 4.836926  13.571829 5.447353e-07 0.0002399857  5.97331796
```

Top Hits list

Functional Enrichment Analysis

After the functional enrichment analysis we can see in which categories the differentially expressed genes are active. This plot was made using the gprofiler website and it shows that there were twenty genes identified. They are active in molecular functions, biological processes and in the cellular components. Our identified genes in the cellular component processes are active in the cell wall and external encapsulating structures. The biological processes happening are pectin catabolic processes and polysaccharide catabolic processes. We have the most active genes in the molecular functions, here pectate lyase activity was registered, as well as carbon-oxygen lyase acting on polysaccharides.

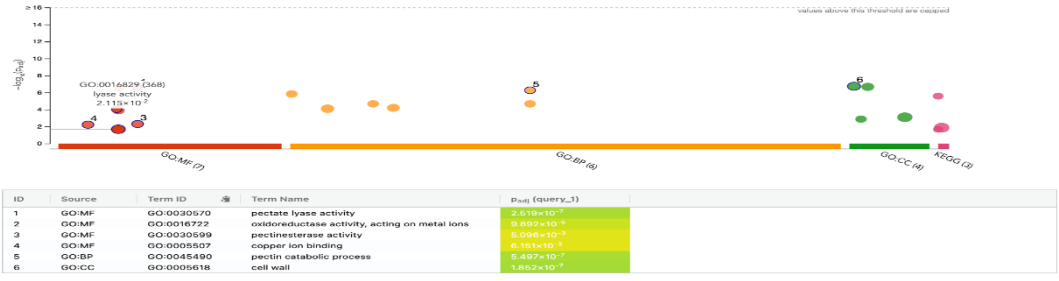
Discussion

Critical Evaluation of the Results

Our sample sequences were downloaded from a trustworthy study, from the NCBI database, to ensure they are high quality. We performed a quality control of those samples and trimming, which made the quality of those samples even higher. The genomes were provided from reliable sources and our analysis was performed with help of our supervisors, which makes us reassured that it was performed correctly. The only problem we met on the way was the evaluation of the results, which might be the issue. We are sure we provided proof that there are differences between a flower and fruit tissue of Solanum Lycopersicoides.

Biological Implications

This study tests the hypothesis that flower tissue of a tomato is different from a fruit tissue, which for some might be surprising, as fruit is a later



ID	Source	Term ID	Term Name	P_{adj} (query_1)
1	GO:MF	GO:0030570	pectate lyase activity	2.619×10^{-7}
2	GO:MF	GO:0016722	oxidoreductase activity, acting on metal ions	9.892×10^{-5}
3	GO:MF	GO:0030599	pectinesterase activity	5.096×10^{-5}
4	GO:MF	GO:0005507	copper ion binding	6.151×10^{-5}
5	GO:BP	GO:0045490	pectin catabolic process	5.497×10^{-7}
6	GO:CC	GO:0005618	cell wall	1.862×10^{-7}

Plot 5 - gprofiler

stage of a flower. The results prove our hypothesis and bring reassurement for other studies, that there are differences in RNA of those tissues.

Limitations and Future Directions

As this study was made by two 3rd semester students after a 3-week practical course, there surely were limitations when it comes to the way the comparison was performed. It would make sense to be recreated by someone with bigger knowledge.

Conclusion

While taking a look at the cluster dendrogram and the PCA results, we can clearly see that our flower samples and fruit samples differ from each other. That result shows us that there are differences between flower and fruit tissue of Solanum Lycopersicoides.

References

Kukurba KR and Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harb Protoc*. 2015;2015 11:951–69. doi:10.1101/pdb.top084970.
https://www.novogene.com/us-en/resources/blog/using-rna-seq-to-understand-plants-and-improve-crop-yield/#
https://en.wikipedia.org/wiki/Tomato
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375638/
https://cyverse-kallisto-tutorial.readthedocs-hosted.com/en/latest/step3.html