Effect of bacterial inoculation on gene expression in *Arabidopsis Thaliana*

Joran Schoorlemmer 1004586

BIF 30806 Group Araformatics

14-12-2021

Abstract

Bacteria can form a dangerous threat to plants. To protect itself, a plant can use many different molecular responses to protect itself from these hazardous pathogens. Why specific molecular pathways are expressed however, is currently not well understood. In this project we look at a paper by (Maier et al., 2021) and use its RNA sequence data of *Arabidopsis Thaliana* plants to rerun their data analysis. We try to find if different bacteria eliciting a strong response constitute to the same pathways being expressed or not. Also, the so called GNSR genes from the paper are looked into, and we try to figure out if these are actually consistently expressed in the different samples. Our results show that similar pathways are expressed in all samples. One of these pathways is the phenylpropanoid pathway, involved in lignification of the cell wall. However, some differences between samples can be found. It is still uncertain why this is the case as these differences are not caused by the phylum or pathogenicity of the bacteria.

Index

Introduction

Plants in nature are often subject of bacterial inoculation. This can be harmless, having the bacteria work together with the plant in symbiosis (Cocking, 2003). Much research on this has been done, because it can improve crop yield greatly and be beneficial to the plant. Bacteria can also form a harmful threat to the organism. They can cause diseases or other forms of crop loss. A lot less is known about the way plants are negatively influenced by its microbiome. It is important to know which genes are important in defense regulatory pathways. This way, we can learn about the function of these genes and try to replicate it (Chezem et al., 2017). We can use this knowledge to better protect the plant, this can lead to bigger crop yields or more resistant crops to droughts or floods. In our case, it can lead to a higher resistance to bacterial diseases.

Some research has been done, for example in a paper from Maier et al (Maier et al., 2021). A study was done on 39 endogenous bacterial strains from *Arabidopsis Thaliana.* This way, they tried to assess host transcriptional and metabolic adaptations to these bacterial encounters. In the paper, they find a general response in most encounters, consisting of 24 genes. They call these genes general non-self-response (GNSR) genes. These genes are all highly differentially expressed. It is proposed that these genes constitute a defense strategy which can be used against diverse strains

and overall contributes to the protection of its host. The GNSR is found to be linked to the tryptophan-derived secondary metabolism.

In this project, we will look at these genes and see if the results obtained by Maier et al. are actually correct. This will be done by recreating their data analysis pipeline using the Maier et al. RNA sequence data. In addition, genes other than the 24 GNSR genes will be looked at. There might be other genes which are also consistently expressed in all samples. This could lead to another general pathway. To structure the research, the following research questions were formulated:

- For bacteria that elicit a strong response, are the same defense pathways triggered or different ones?
- Do these defense pathways differ from the GNSR genes pathway? If so, in what way?

As can be seen in the research question, only bacteria that elicit a strong response are looked at. Bacteria eliciting a strong response might elicit different pathways as lowly expressed pahtways. While some pathways might only need little expression of its genes to perform its function, others might need a high expression for its proteins to perform its function correctly (Karasov et al., 2017), (Kwon et al., 2020).

The results could also differ from the results in the paper of Maier et al. because a subset of the samples is taken in account. Other reasons for creating this subset are that these strong responses will yield the clearest and most differentially expressed results. Also, the sample size is decreased a lot which speeds up the research process by decreasing running time on the cluster.

Design

The entire data analysis pipeline had to be recreated. This process was divided into three main parts.

1. Mapping and annotation
2. Counting and Differential expression analysis
3. Clustering and GO/pathway enrichment

At first, the idea was to include all three parts into one big snakemake file. In the end, the three parts were all in separate steps. It will be discussed later why this combination of steps was more challenging than expected. The full pipeline is shown below in Figure 1.

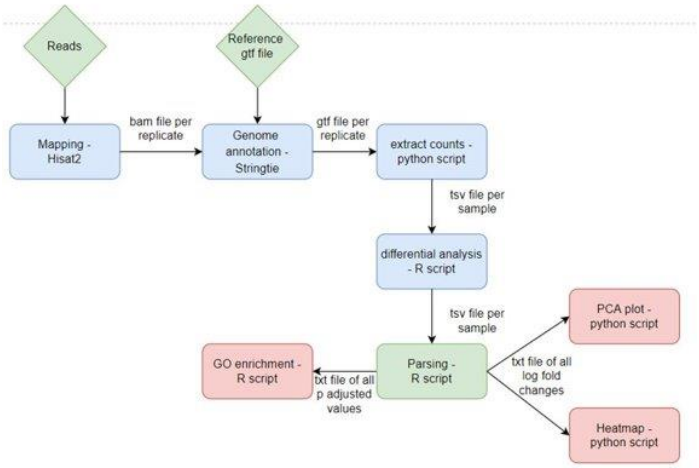


Figure 1: Data analysis pipeline

To speed up the coding process, a test data set was created. This way, scripts and tools could be validated while only needing little processing capacity. This data set consisted of two samples with 2 replicates per sample. The samples were Leaf51 and the axenic control from the paper.

Mapping and annotation

RNA sequence data was provided by Maier et al. To see from which genes these RNA sequences are transcribed, it is necessary to map these reads to the genome of *A. Thaliana.* In this project, HISAT2, version 2.1.1, was used for the mapping(Kim et al., 2019). In the Maier et al. paper, RSEM was used. This was changed in this project due to the speed of RSEM. It was expected that RSEM would take at least 20 times as long as HISAT2 which wasn't a possibility in the short timescale of the project. HISAT2 was chosen due to its ability to also map introns and jump over these splice sites. This could lead to more reads being mapped opposed to e.g. bowtie. TAIR10 was used as a reference genome, this is the same as in the paper.

No quality check of the reads was done and as such no reads were trimmed before mapping. This choice was made based on prior knowledge, both from the paper and the teaching staff of BIF30806. They found little low quality reads and this was also confirmed by other students doing similar projects.

For annotation, Stringtie 1.3.2d (Pertea et al., 2015) was used. While doing the differential expression analysis, we found multiple transcripts being annotated to one single gene. This resulted in overall lower count values per transcript as these were divided over multiple imaginary transcripts by stringtie. To exclude multiple transcripts mapping to the same gene, the -e parameter was used while running stringtie. This option makes sure stringtie only returns the expression values of the transcripts given in the reference genome.

To validate the mapping, the bam and bam index file were loaded in iGV to the TAIR10 genome. The alignments were checked by eye. Reads from replicates were found to be correctly mapped to the genes, an example is shown in Figure 2.

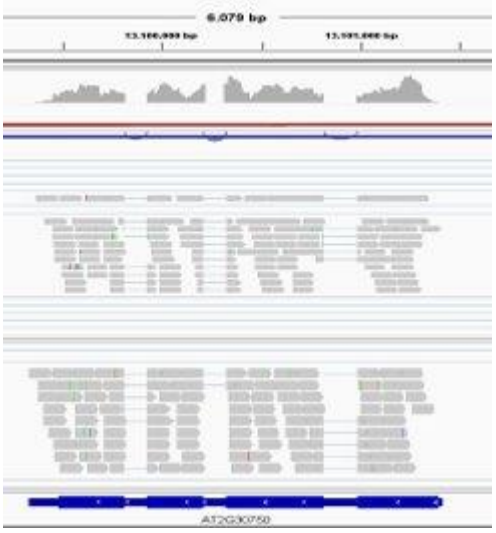


Figure 2: reads of two replicates of sample Leaf51 mapped to the TAIR10 genome by Hisat2. The scales differ a bit for the two replicates but they show a similar alignment.

Counting and differential expression analysis

To translate the .gtf annotation file contents into actual counts per transcript, Equation 1 was used.

Equation 1: formula to calculate counts with transcript_length and read length in basepairs.

$$counts\ per\ transcript = coverage * transcript\_length\ / read\_length$$

A basic python script was used to parse the contents of the gtf files, using the coverage and transcript length given for each transcript. The counts were calculated and written to a tsv file. The annotation step returned all genes in the reference TAIR10 genome, whether they were expressed or not. Because of this, a cutoff coverage was used of 0.01. Only transcripts with a coverage higher than 0.01 were taken in consideration. This differs from the paper as they use a cutoff of 0.5 Counts per Million. In this project, the coverage was used instead of CPM to improve the understandability of the script. 0.01 was taken as a cutoff as this resulted to a CPM of 0.5 for transcripts of an average length.

These count values were used as input for the actual differential expression analysis using DESeq2 (Love et al., 2014). In the paper by Maier et al., EdgeR was used. However, due to the easier access to DESeq2 and some technical difficulties with installing EdgeR on the server, we used DESeq2. These packages uses very similar techniques. First, the counts are normalized. In EdgeR, this is done by a trimmed mean of M values method while we used a log transformation to normalize the counts. Now, the gene wise dispersion were estimated. These dispersion values were used to calculate a standard error (SE). This SE, together with the base mean of a generalized linear model was used in a binomial Wald test to determine the log2fold changes and a p-value. The p-values were corrected for a high FDR with the same Benjamini-Hochberg method as was used in the paper. In the test, a log2fold change of 1 was taken as null hypothesis while the alternative hypothesis was that a gene has to have a bigger absolute log2fold change than 1. This is adjusted to prevent artificially low p-values as opposed to a null hypothesis of 0.

For each sample, these fold changes and p-values were calculated. To make further analysis easier, a simple python script was written to parse these csv files into one big csv file for all fold changes and one csv file for all p-values of all samples.

The results were validated by counting some reads ourselves and checking if these compare to the results from our scripts.

Clustering and GO/pathway enrichment

In the Maier et al. paper, genes were clustered using Ward's method. This gives a precalculated dendrogram which was not feasible in our python script. In this project complete linkage was used as an alternative to ward's method. The distance between samples and genes was calculated by Pearson correlation. The result from the clustering was visualized in a heatmap. A principal component analysis was also performed on the data using the log2fold changes. For the clustering, a subset of 21 genes was used to decrease workload. These 21 genes all came from the GNSR genes mentioned in the paper.

The GO enrichment analysis was done using topGO (Alexa & Rahnenführer, 2016) as opposed to AgriGO2v2 which was used in the paper. This adjustment was made due to similar reasons as the DESeq2 choice, AgriGO2v2 yielded to many technical difficulties. TopGO was used instead of other, more general methods like performing a manual fisher test because topGO already comes with a premade list of all GO terms involved in biological processes. TopGO performed a fisher test on the p-values obtained in the DE analysis. Afterwards, these p-values were corrected with the same

Bejamini-Yekutieli (BY) method as was used in the paper. These p-values were compared to the GO terms in topGO.

The pathway enrichment was done using KEGG, similar statistical operations were performed as in the GO enrichment. For these analyses, all differentially expressed genes were used.

The results of the pathway enrichment were validated by performing an analysis using KOBAS webtool (Bu et al., 2021). Again, a Fisher's exact test was used with an BY FDR correction.

Results

The first real results are obtained in the differential expression step. Differentially expressed genes were found for all seven samples, including the reference. These are shown in Figure 3 and Figure 3.

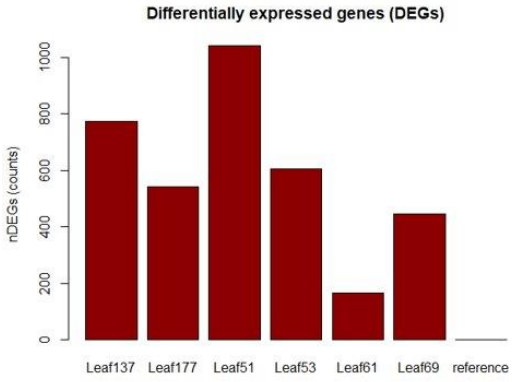


Figure 4: Differentially expressed genes (DEGs) found in all samples in this project.

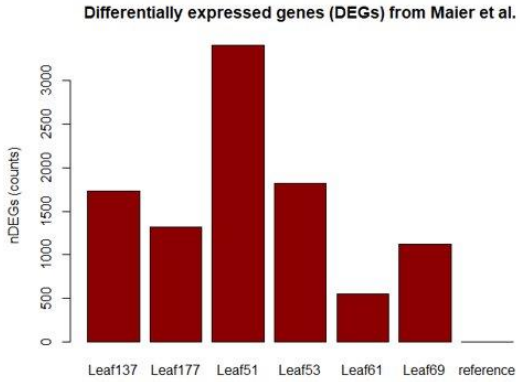


Figure 3: Differentially expressed genes (DEGs) found in all samples in Maier et al. paper.

As can be seen in these figures, the number of found DEGs is approximately 1/3 of the amount found in the Maier et al. paper. It will be discussed later why this is the case.

Also, some MA plots were made for the seven samples. As expected, the reference showed no differential expression at all. An example is shown in Figure 5.

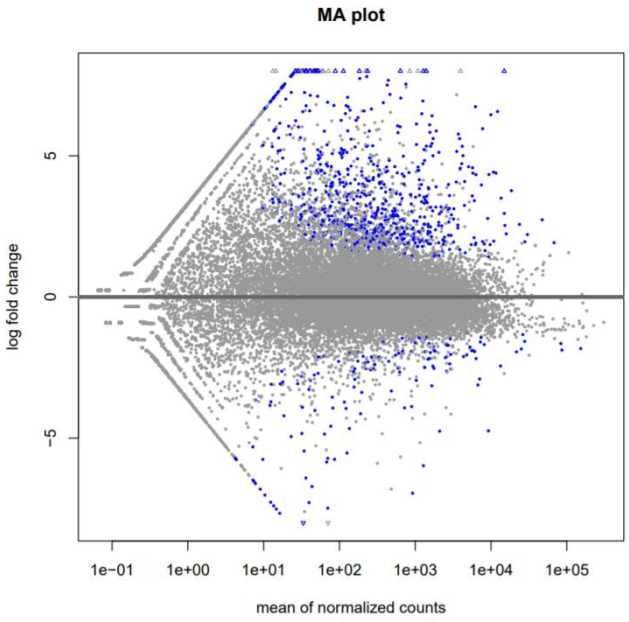


Figure 5: MA plot of expressed genes in sample Leaf51. Genes with a significant expression (p_{adj}< 0.01) are shown in blue.

As can be seen in the figure, many genes are not significantly expressed. Only few genes are marked in blue. Another notable observation is that most genes, significant or not, are upregulated. This is consistent in all samples.

The resulting heatmap from the clustering is shown in Figure 6. The big blue bar on the left is the reference which of course shows no fold change at all. These genes show very similar expression in most samples. Only the third column, Leaf61, shows a slightly different expression pattern for some genes. Again, most genes are upregulated which can be seen from the predominantly red boxes.

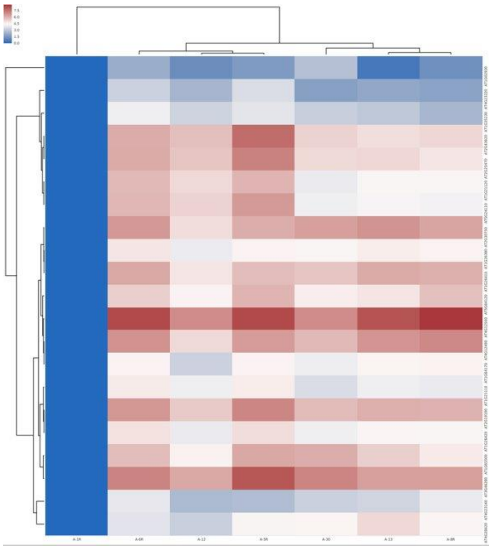


Figure 6: Heatmap of DEGs, clustered by complete linkage based on pearson correlation distance. Samples are shown on the X axis and genes on the Y axis.

The results from the GO enrichment analysis showed similar results for all bacteria too. A visualization of the GO terms in Leaf137 is shown in Figure 7.

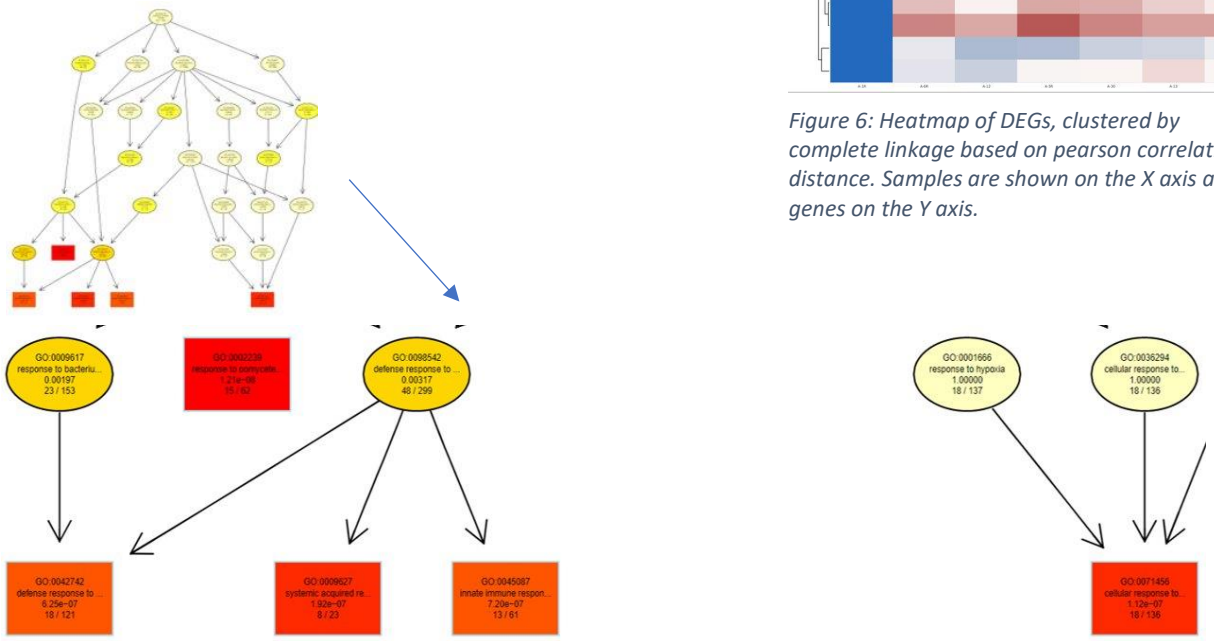
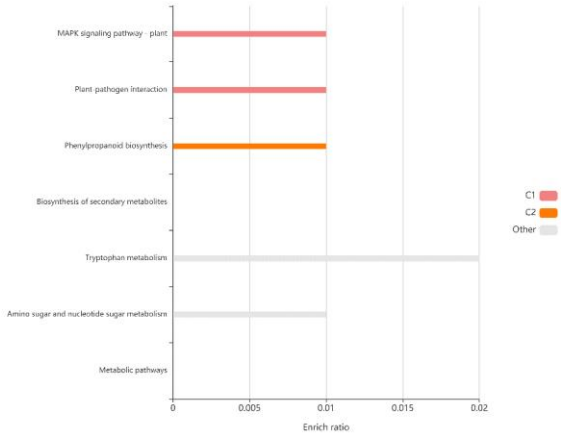


Figure 7: Significant found GO terms in sample Leaf137.

As can be seen in the figure, only few GO terms are actually significant. Some notable ones are GO 0042742 which is a defense response to bacterium term and GO 0071456 which is a cellular response to hypoxia term. Hypoxia is a situation in an organism in which a shortage of oxygen is present.

The KEGG pathway analysis resulted in only one pathway being upregulated in all bacteria. This is the Phenylpropanoid pathway. 1 gene in this pathway is also found in the 24 GNSR genes mentioned in the Maier et al. paper. Other pathways were found but not consistently in all samples. Some examples are Amino sugar and nucleotide sugar metabolism in Leaf137, Leaf69 and Leaf61 and Glycosphingolipid biosynthesis in Leaf53 and Leaf177.

Some results of the pathway analysis using the online KOBAS tool are shown in Figure 7 and Figure 8.



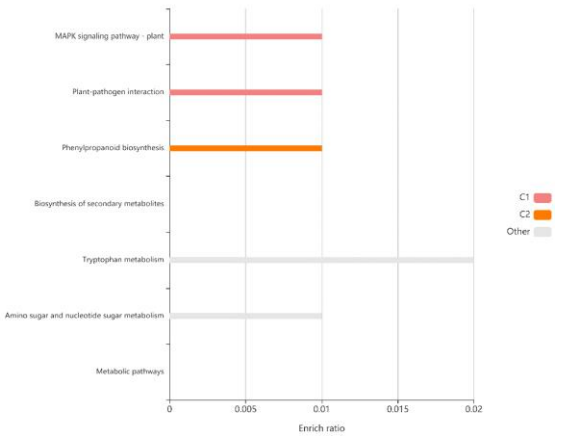


Figure 8: pathway analysis using KOBAS for sample Leaf51 Figure 9: pathway analysis using KOBAS for sample Leaf61

As can be seen in the figures, the KOBAS pathway analysis is very similar for all samples. The same phenylpropanoid pathway is found as in the KEGG pathway analysis. Also, the tryptophan metabolism is found to be expressed. This is similar to the paper by Maier et al., stating that the GNSR is linked to the tryptophan-derived secondary metabolism.

Discussion & conclusion

The main research questions of this project are:

- For bacteria that elicit a strong response, are the same defense pathways triggered or different ones?
- Do these defense pathways differ from the GNSR genes pathway? If so, in what way?

In the used bacteria samples, multiple defense pathways were triggered. However, the results differed for the two main techniques used. In the KEGG analysis, only one pathway was found in all samples with multiple differing pathways between the samples. There seems to be no relation between the phylum of the sample and the expressed pathways. This can be concluded because all samples are from different phyla (Maier et al., 2021). This is not expected, because bacteria in the same phylum usually share more similarity and will use similar attack strategies.

In the KOBAS analysis, all found pathways were very much the same for all samples as can be seen in Figure 8. This is a big difference from the KEGG analysis. However, this does fit better with the idea of a GNSR as proposed by Maier et al. The reason for this difference in the two analyses could be a technical one. In the KEGG analysis, only genes starting with AT1G were taken in account. This way, many genes were left out. This could lead to less significant pathways being found. This also explains why only one pathway was expressed in all six samples, which is not expected according to Maier et al.

This result is also supported by the clustering, in which all samples show similar expression values for all genes. However, in the clustering only 21 genes were used to decrease the amount of calculations needed. All these 21 genes are from the set of GNSR genes so it is expected that they behave the same in all samples. In future research, the clustering can be expanded by adding many more genes. This could lead to more interesting results, like a subset of genes being expressed in a specific bacterial phylum.

To answer the second research question, the KEGG analysis can still be used. If information is left out but pathways are still significantly expressed, we know that they would have been expressed for sure in the full analysis. Here, it has been found that only one gene corresponds to the GNSR genes from the paper. This suggests other pathways are triggered in the bacteria eliciting a strong response. This could happen because these pathways are so highly expressed that the GNSR pathway is not needed anymore. The KOBAS analysis gives a clear indication that the tryptohan metabolism pathway is expressed as is proposed by Maier et al. The full data set gives a more similar answer to the paper. In this analysis however, the individual genes making up the pathway were not looked into.

Another point to address is the low amounts of found DEGs. In this project, on average 1/3 of the DEGs were found as opposed to the paper by Maier et al. The consistency of the found DEGs being 1/3 of the paper DEGs indicates that there is a structural flaw in the paper or this project. This could be due to multiple reasons. The samples all consisted of 5 replicates. If one of these replicates contains a big outlier count value for a gene, DESeq will give a NA value for the p value as it can't do a reasonable test on this data. While the data was log transformed, still many genes yielded NA values after the DE analysis. This can be resolved by slicing out the outlier replicates. Another reason could be the cutoff coverage for the DE analysis. In this project, only transcripts with a coverage > 0.01 were taken in account as opposed to a CPM > 0.5 as was used by Maier et al. This can lead to too little genes being taken in account, especially for the short transcripts.

Combining all three data analysis steps turned out to be harder as expected. Files often needed slight formatting which was more work to automate in a snakemake file than to just e.g. manually move all samples to a different folder. For this reason, only the mapping and annotation were run in a snakemake file while the others where just run manually. This was automated a bit by listing filenames in a text file and reading the data using this text file. This way, all samples could be ran by one single command instead of running the code for each individual sample. This can be automated a lot more by adding these steps to the snakemake file.

In the future, this research can be improved a lot. Many more genes can be taken in account in the clustering and pathway analyses. Also, all samples from the paper can be used. This will probably give a more general conclusion than found in this project. Also, the difference between the mapping with RSEM and Hisat2 can be looked into, as Hisat2 actively skips introns while this is not necessarily the case for RSEM.

All in all, we can conclude that multiple pathways are expressed in plants inoculated with bacteria eliciting a strong response. However, the response of *A. Thaliana* is very similar when inoculated by different bacteria. The general non-self response proposed by Maier et al. is found in all samples but another pathway is also consistently found. This is the phenylpropanoid pathway. This pathway is involved in the lignification of cell walls (Dixon et al., 2002)so it makes sense that this pathway is expressed while the plant is under attack of hazardous pathogens.

Contributions

Me and Ward Koehler worked together on the counting and differential expression step. I did the most of the scripting, resulting in a python script extracting count values from gtf files. A R script doing the DE analysis using DESeq2. Another python script parsing these DE results in the correct format and another R script to visualize some of the DE analysis results. Meanwhile Ward checked my scripts and did background research on parameter settings for e.g. DEseq2. Also, maintaining the groupfolder on the server and the gitlab project were done by me.

References

Alexa, A., & Rahnenführer, J. (2016). *Gene set enrichment analysis with topGO*. http://www.mpi-sb.mpg.de/~alexa

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., Wu, Y., Zhao, L., Liu, J., Guo, J., Fang, S., Cao, W., Yi, L., Zhao, Y., & Kong, L. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Research*, *49*(W1), W317–W325. https://doi.org/10.1093/NAR/GKAB447

Chezem, W. R., Memon, A., Li, F. S., Weng, J. K., & Clay, N. K. (2017). SG2-Type R2R3-MYB Transcription Factor MYB15 Controls Defense-Induced Lignification and Basal Immunity in Arabidopsis. *The Plant Cell*, *29*(8), 1907–1926. https://doi.org/10.1105/TPC.16.00954

Cocking, E. C. (2003). Endophytic colonization of plant roots by nitrogen-fixing bacteria. In *Plant and Soil* (Vol. 252).

Karasov, T. L., Chae, E., Herman, J. J., & Bergelson, J. (2017). Mechanisms to Mitigate the Trade-Off between Growth and Defense. *The Plant Cell*, *29*(4), 666–680. https://doi.org/10.1105/TPC.16.00931

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology 2019 37:8*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kwon, J., Bakhoum, S. F., & Kettering, S. (2020). The Cytosolic DNA-Sensing cGAS-STING Pathway in Cancer. *Aacrjournals.Org Cancer Discov*, *10*, 26–39. https://doi.org/10.1158/2159-8290.CD-19-0761

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), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Maier, B. A., Kiefer, P., Field, C. M., Hemmerle, L., Bortfeld-Miller, M., Emmenegger, B., Schäfer, M., Pfeilmeier, S., Sunagawa, S., Vogel, C. M., & Vorholt, J. A. (2021). A general non-self response as part of plant immunity. *Nature Plants 2021 7:5*, *7*(5), 696–705. https://doi.org/10.1038/s41477-021-00913-1

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology 2015 33:3*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

The phenylpropanoid pathway and plant defence-a genomics perspective. (2002). *MOLECULAR PLANT PATHOLOGY*, *3*(5), 371–390.