TransciptMetaAnalyst User Manual

Eslam M. Abdel-Salam*

TranscriptMetaAnalyst (TMA) is a graphical user interface to meta-analyze RNA-Seq transcriptomics data. It is designed to help biologists with less programming skills to perform meta-analysis of several transcriptomic RNA-Seq-generated data sets in order to get robust results and new insights into the data that was generated by their own studies or the data available online.

Table of contents

*Plant Molecular Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Munich, Germany

1 Introduction

Meta-analysis of RNA-Seq transcriptomics data has become an essential tool in modern biology to identify differentially expressed genes (DEGs) and pathways across multiple studies[1,2]. However, performing meta-analysis can be a challenging task, especially for biologists with limited programming skills. To address this issue, we have developed the TransciptMetaAnalyst (TMA) app, a graphical user interface (GUI) designed to facilitate the meta-analysis of RNA-Seq transcriptomics data.

TMA allows users to easily combine and analyze multiple RNA-Seq datasets, providing a robust and reliable way to identify differentially expressed genes and pathways. The app is built on top of the popular R programming language and utilizes the metaRNASeq package[3] to perform meta-analysis. With TMA, users can upload their own RNA-Seq data or use publicly available datasets along with a simple experimental design file, and run the whole analysis with a single click to get publication-ready figures and tables.

TMA is particularly useful for researchers who want to:

- Identify differentially expressed genes and pathways across multiple studies
- Combine data from different studies to increase statistical power
- Visualize the results of the meta-analysis using different methods, e.g., heatmaps and volcano plots

By providing a user-friendly interface to perform meta-analysis, TMA aims to facilitate the discovery of new insights and robust results in transcriptomics research.

2 Installation

The app could be installed locally via GitHub using the following commands (please note that `BiocManager` should be installed before installignt he app):

```
# Check if devtools and BiocManager are installed, and install them if not
if (!requireNamespace("devtools", quietly = TRUE)) {
  install.packages("devtools")
}
if (!requireNamespace("BiocManager", quietly = TRUE)) {
  install.packages("BiocManager")
}
# Now you can use devtools to install the ShinyApp
devtools::install_github("Eabdelsalam/TranscriptMetaAnalyst")
# Finally, run the app
shiny::shinyAppDir(system.file(package = "TranscriptMetaAnalyst"))
```

The first time of running the app might take longer to launch the GUI as all the dependency needed packages will be checked and the missed ones will be installed.

3 Quick Start

To start the RNA-Seq meta analysis, please click on 'RNA-Seq Meta Analysis' tab and perform the following steps:

- After cleaning the raw RNA-Seq read files (either downloaded from public databases of generated via your own experiments) of each sample, align them individually to the studied organism's reference genome.
- Count the mapped reads to the reference genomic features of interest and combine all the counts in one file. Keeping genomic features as rows and the samples that will be included in the analysis as columns. To produce the final raw (unnormalized) counts comma-separated file. ***Note that the counts file should have a 'Geneid' column***.
- Upload the counts file produced by any feature counting tool, e.g., featureCounts or htseq-count, as explained above to the raw counts placeholder. ***Please refer to example counts file via the provided download link***.
- Upload the experimental design file. ***Please refer to the example file via the provided download link***. This file should contain at least 3 columns, namely 'sample' refers to the sample names used as columns in the counts file, 'study' refers to the study to which the sample belong, 'condition' refers to the treatment or control conditions (only two conditions allowed for comparison).
- Hit ***Start Analysis!*** button and wait for the results to appear.
- To perform further **Functional Annotation** of the DEGs identified via the meta-analysis, the functional annotation box in the bottom-left corner of the page could be used. The user needs to choose the organism before performing this analysis. Please also consider that this analysis needs an active internet connection to be performed and it might take a while. So, please be patient *_*.

4 App Layout

4.1 Home Tab

After running TMA, the layout shown in Figure 1 will open in a separate window. The top left corner holds the app name. The left panel holds different tabs that form the core of the app and its functionality. The three bar symbol besides the app name could be used to toggle on/off the side panel. The **Home** tab opens by default upon running the app and contains brief information about the app and its components.

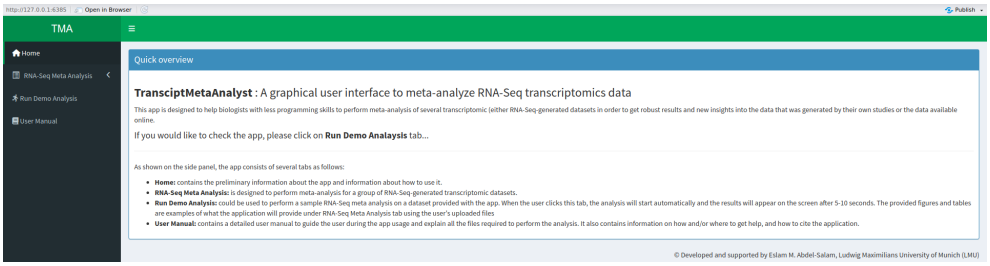


Figure 1: Layout of the TranscriptMetaAnalyst app. The app opens the home tab by default showing brief information about the app and how to use it.

4.2 RNA-Seq Meta Analysis

This tab contains all the needed information about the required steps to perform a meta-analysis of RNA-Seq data using TMA. As shown in Figure 2, the tab shows the *Quick Start* steps to run the analysis. On the left panel, there are two placeholders for the required input files (discussed below). Moreover, there are two links that allow the user to download example (template) files that will guide the user in preparing the raw counts and the experimental design files or could be used as templates to copy and paste their own counts in.

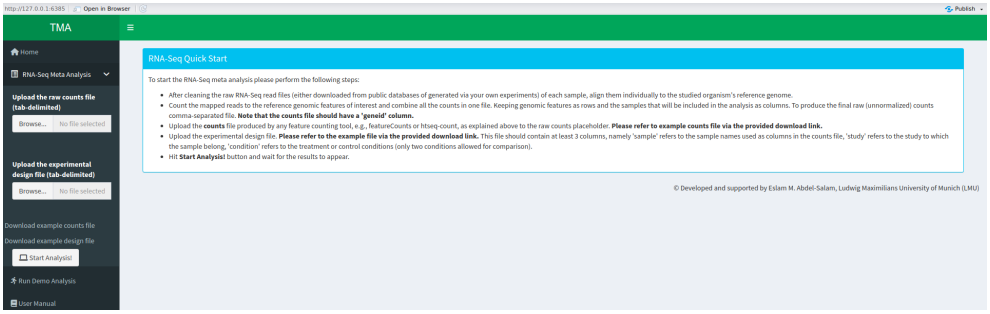


Figure 2: RNA-Seq Meta Analysis tab. This is the main tab of the TMA functionality that allow the user to analyze their own data.

4.2.1 Input Files

To perform a meta-analysis of RNA-Seq data using TMA, the user need two input files:

1. The ***raw counts file*** that is a tab-delimited file contains the raw counts of all the experiments to be analyzed, where each row represents a single gene and each column represents all counts from a single sample. Figure 3 shows an example of the raw counts file.

Geneid	Chr	Start	End	Strand	Length	gene	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	Sample11	Sample12
slr0612	CHP	811	1494	+	684	NA	587	705	591	491	94	208	221	244	3	3	5	0
slr0613	CHP	1577	2098	+	522	NA	2082	2900	2790	2056	251	585	664	655	3	3	10	0
sll0558	CHP	2172	2873	-	702	NA	2163	2923	3269	2880	272	738	908	603	6	9	12	5
sll1214	CHP	3192	4268	-	1077	acsF	17056	17716	33084	32404	2397	10782	10282	7268	97	107	179	53
sll1213	CHP	4451	5389	-	939	NA	4367	4975	5879	4906	307	1160	1174	792	20	15	30	12
sll1212	CHP	5534	6622	-	1089	gmd	5676	6628	6871	6097	589	1696	1665	1298	24	32	30	15
slr1311	CHP	7229	8311	+	1083	psbA	546817	448849	263259	273388	82695	524857	713436	464928	2891	2761	18421	5365
slr1312	CHP	8492	10471	+	1980	speA	3266	3962	5555	4568	514	1441	2165	1847	16	7	25	9
sll1209	CHP	10622	12631	-	2010	ligA	819	904	1133	1027	309	675	517	788	5	8	3	2
slr1315	CHP	12782	13363	+	582	NA	309	409	500	415	116	283	233	313	3	4	14	4
slr1316	CHP	13498	14529	+	1032	NA	294	445	470	379	162	303	126	330	0	6	5	1
slr1317	CHP	14626	15675	+	1050	NA	335	383	350	321	132	236	118	254	0	3	2	1
slr1318	CHP	15937	16743	+	807	NA	488	486	568	535	131	392	126	273	2	0	2	0
slr1319	CHP	16740	17687	+	948	NA	405	485	504	461	145	311	195	335	1	2	1	2
sll1206	CHP	17853	20408	-	2556	NA	388	500	577	494	346	573	221	646	1	3	4	4
sll1205	CHP	20615	21595	-	981	NA	477	591	589	582	170	409	233	372	0	4	3	1
sll1204	CHP	21838	23121	-	1284	NA	78	126	83	42	133	401	114	308	0	0	0	0
sll1203	CHP	23139	24104	-	966	NA	53	60	60	58	96	305	77	212	0	0	1	0
sll1202	CHP	24101	25003	-	903	NA	33	53	32	28	56	220	62	98	0	0	0	0
sll1409	CHP	25266	27839	-	2574	NA	144	180	159	120	443	1064	366	778	2	3	3	0
sll1408	CHP	27990	29030	-	1041	NA	101	137	106	83	61	330	158	180	1	1	2	0
sll1407	CHP	29098	29895	-	798	NA	60	73	69	50	99	391	82	181	0	0	0	0
sll1406	CHP	29968	32454	-	2487	NA	416	422	389	301	324	691	381	695	4	5	9	1
sll1405	CHP	32524	32928	-	405	NA	69	56	69	52	54	73	55	76	0	0	1	1
sll1404	CHP	32912	33544	-	633	NA	166	209	220	232	58	158	61	114	1	2	0	0
slr1484	CHP	33891	35489	+	1599	NA	484	486	710	619	343	562	233	655	2	3	4	2
slr1485	CHP	35566	36615	+	1050	NA	562	978	729	411	227	423	158	419	1	0	4	0
slr1488	CHP	36757	38538	+	1782	NA	280	404	368	300	310	561	152	620	2	0	3	2

Figure 3: An example of the raw counts file needed for the analysis. Note that the file contains a 'Geneid' column.

2. The ***experimental design*** file that should at least contain 3 columns, i.e., sample, study, and condition. An example of this file is shown in Figure 4. The study column should contain records for all the count columns in the *raw_counts_file*. The condition column should contain only two unique conditions to be compared via the meta-analysis.

sample	study	condition
Sample1	A	Treated
Sample2	A	Treated
Sample3	A	Cont
Sample4	A	Cont
Sample5	B	Cont
Sample6	B	Cont
Sample7	B	Treated
Sample8	B	Treated
Sample9	C	Cont
Sample10	C	Cont
Sample11	C	Treated
Sample12	C	Treated

Figure 4: An example of the experimental design file. This example file contains the three required columns only.

4.3 Run Demo Analysis

The Demo Analysis tab (Figure 5) is designed to provide users with a hands-on experience of how the TransciptMetaAnalyst app works. By clicking the ***"Run Demo Analysis"*** button, users can run a demo analysis using simulated data of *Synechocystis* sp. based on our previously published meta-analysis[1]. After clicking, a message showing that the demo analysis is running will appear. After a while, the message will disappear and the results

will be visible to the user. All the outputs of the TMA will be discussed later in the outputs section. This demo analysis will generate the entire results, including all the meta-analysis outputs and figures, which will be discussed in detail later in this manual. This feature allows users to test the app's functionality, understand the analysis workflow, and get an overview of the app's capabilities before working with their own data. By running the demo analysis, users can also familiarize themselves with the app's user interface and learn how to navigate through the different tabs and sections.

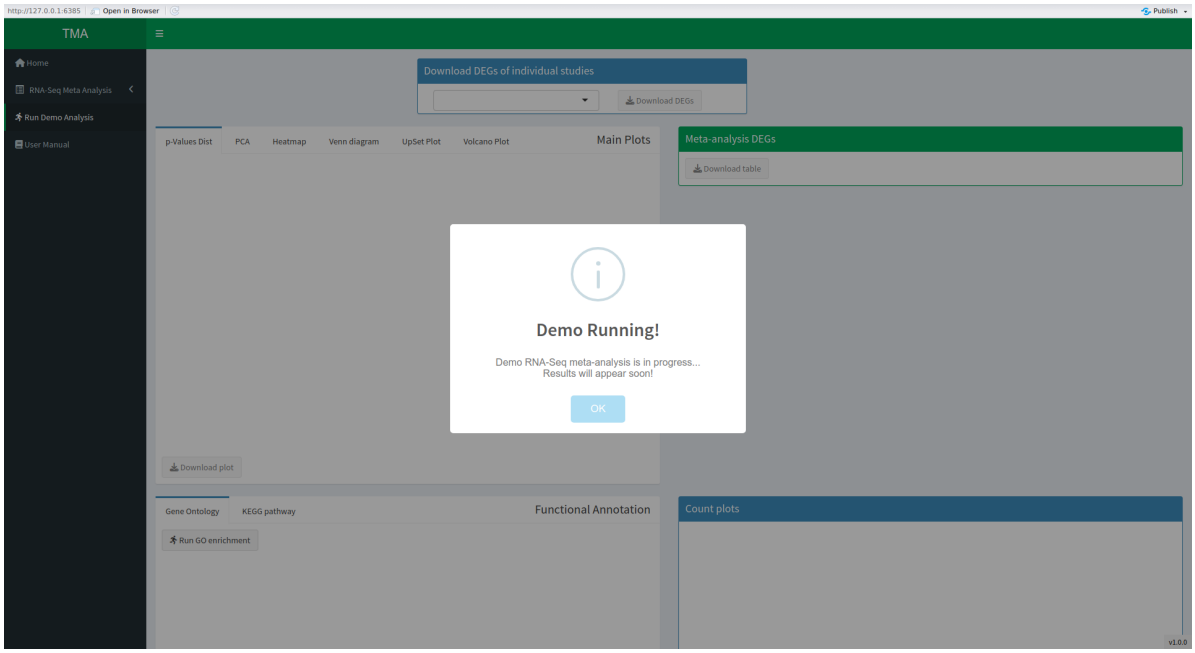


Figure 5: TMA provides a Run Demo Analysis tab that allows the users to test the app. Clicking on this tab will trigger the analysis and the results will appear in a short notice.

4.4 User Manual

The last tab on the left panel is the *User Manual* tab provides the user with a link to access the current detailed guide of TMA. It also provides the user with information regarding contacting the creators of the app and how to cite the app in scientific publications (Figure 6).

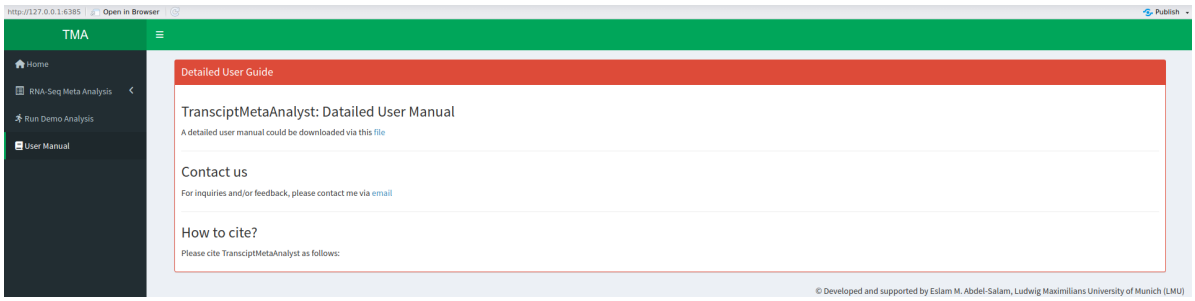


Figure 6: The User Manual tab of TMA. This tab provides the user with links to download the current detailed guide and to contact the app's creators along with citation information.

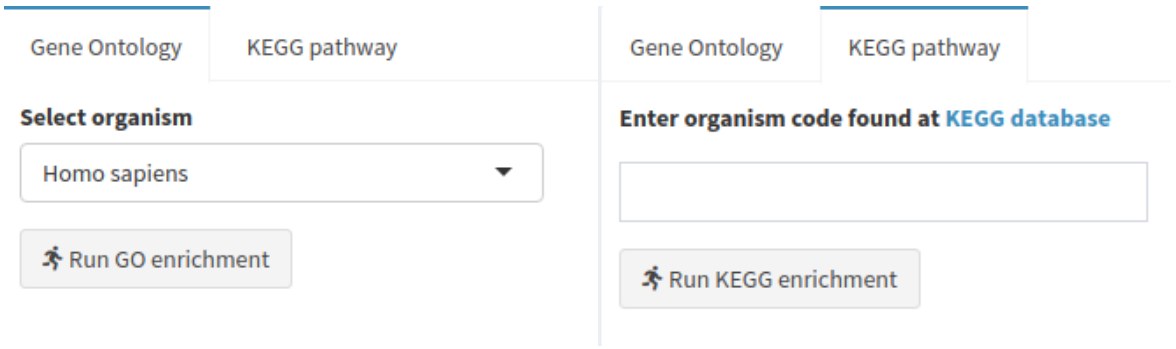
5 Analysis Pipeline

After uploading the required files, several checks will be performed to make sure that the files are in compliance with the analysis requirements. The user will be notified about any problems in the files and asked to fix them before re-run the analysis.

If the files passed the pre-check, the analysis will start automatically and a load screen will appear until the results are prepared and shown to the user. For those who need to know what is going behind the scene, DEGs are identified in each individual study separately using DESeq2 package[4]. The raw p-values are then combined using Fisher's method[5] via functions from the metaRNASeq package[3] to identify DEGs across all the included studies. As a validation step, all genes with conflicts (those show up-regulation in one study and down-regulation in another) are removed from the final list of meta-analysis DEGs.

If the user would like to perform further functional annotation of the identified DEGs, Gene Ontology (GO) and KEGG pathway enrichment analysis could be performed. In the case of GO, the user needs to select the studied organism from the provided drop-down menu (Figure 7a) and run the analysis. This analysis is based on the PANTHER Classification System v17.0[6] and performed utilizing the `rbioapi` package. The KEGG pathway enrichment analysis is performed via the clusterProfiler package[7]. The user also needs to write down the 3-letter code of the studies organism that could be obtained from the KEGG Organisms Database before running the analysis (Figure 7b).

Finally, to produce the different outputs (to be discussed in the coming section), several in-house scripts along with built-in functions and/or other packages in R are used. The outputs are grouped and shown in different boxes. All the outputs could be downloaded as PDF files for the figures or CSV files for the tables.



(a) GO Enrichment

(b) KEGG Pathway Enrichment

Figure 7: Gene Ontology (GO) and KEGG Pathway enrichment analysis in TMA. The user should (a) select the organism from the drop-down menu before running the GO analysis or (b) type the 3-letter code of the studied organism before running the KEGG Pathway analysis.

6 Outputs

6.1 Overview & Informaiton Boxes

TMA provides the user with 7 main output boxes (Figure 8). At the top section of the results screen, two information boxes could be found on the left and right sides showing the number of the studies included in the analysis and the number of DEGs identified via the meta-analysis, respectively. In the middle, there is a box that allow the user to download a list of DEGs that were identified in each individual study included in the analysis.

Below the information boxes the results screen is divided into 4 main section. The top left box contains the main plots (to be discussed later), the top right box contains a table of all DEGs identified in the meta-analysis with their statistical parameters. This table allows search functions and different order methods. It could be also downloaded as CSV file for further analysis. Clicking on any row of the meta-analysis DEGs table with lead to update in the Counts plots box (the bottom right) to show plots of the normalized counts of the selected gene in the table across all the studies included in the meta-analysis. These plots could also be downloaded as PDF file. The bottom left box contains the GO and KEGG Pathway enrichment analysis and results (to be discussed later).

6.2 Main Plots

TMA generates several main plots to provide the user with an overview of the analyzed dataset and the identified DEGs.

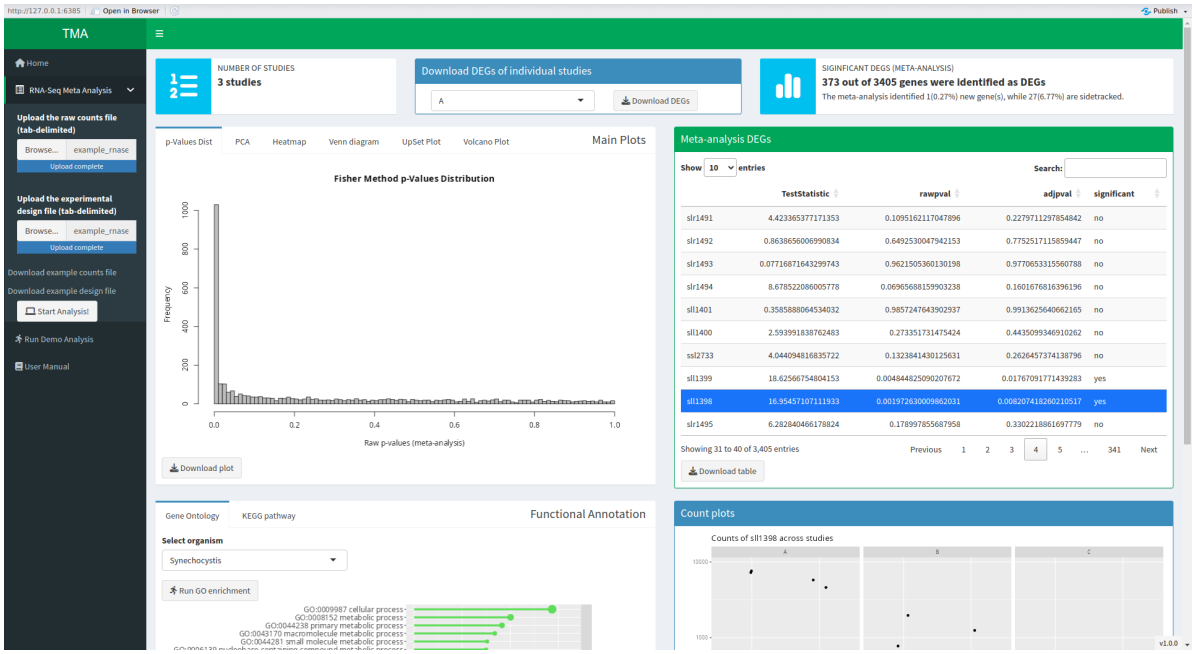


Figure 8: An overview of the outputs generated by TMA.

1. The p-values distribution plot (Figure 9) shows the distribution of all raw p-values included in the meta-analysis which allow the user to check the distribution of the p-values. The applied combination method of p-values method assume a uniform distribution under null hypothesis. Therefore, no peaks other than the one near zero should be appear in the distribution.

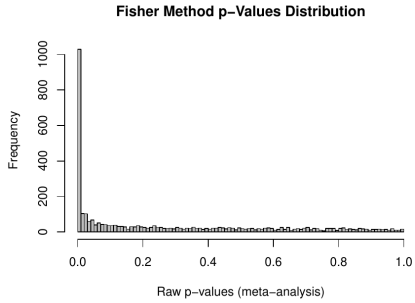


Figure 9: The p-values distribution histogram generated by TMA.

2. The Principal Component Analysis (PCA) is performed via DESeq2 package[4]. The output figure (Figure 10) could be used to identify the relationship between different samples and/or biological replicates, in addition to identify potential outliers and exclude them from the analysis.

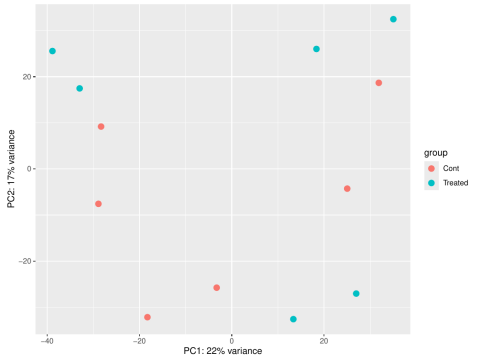


Figure 10: An example of the PCA plot generated by TMA.

3. A heatmap (Figure 11) of the normalized counts of the DEGs identified in the meta-analysis is generated. In TMA GUI, the heatmap is interactive and build using `heatmaply` package[8]. The user could download the figure (not interactive) or the data used to draw the heatmap for further analysis.

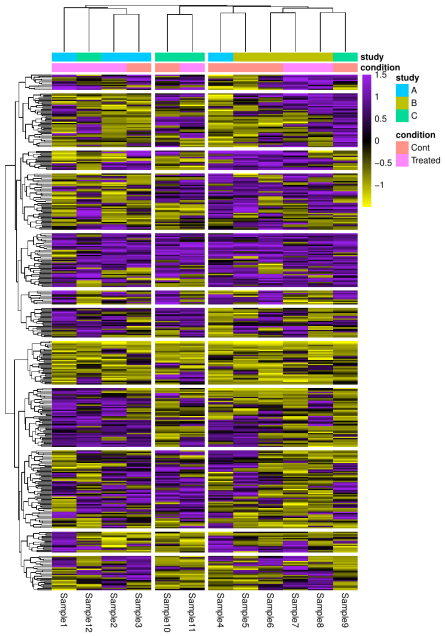
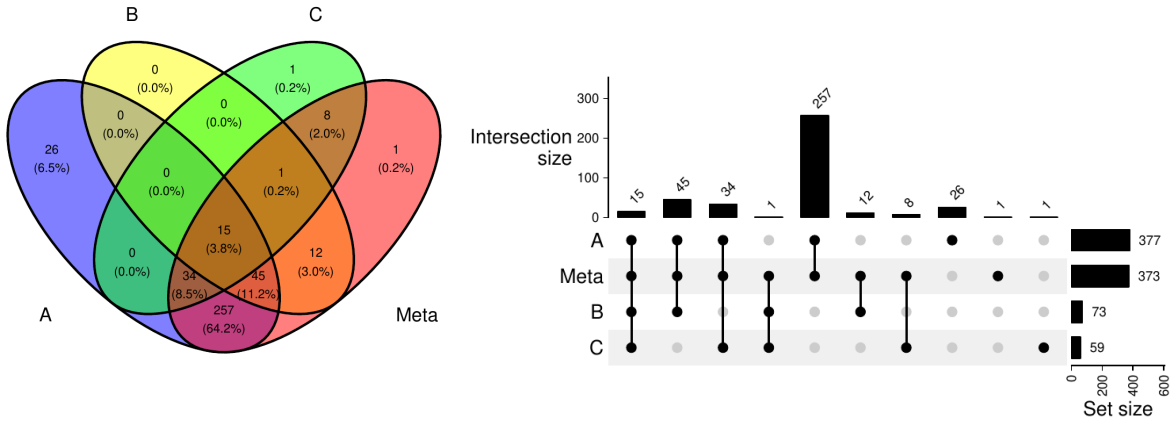


Figure 11: The heatmap generated via TMA could be downloaded as a non-interactive version.

4. A Venn diagram (Figure 12a) and an UpSet plot (Figure 12b) showing the intersection of identified DEGs across the analyzed studies and the meta-analysis is generated. If the number of studies included in the meta-analysis are more then 3 studies, the Venn diagram is disabled and only UpSet plot is generated.



(a) Venn Diagram (b) UpSet Plot

Figure 12: The intersection of identified DEGs is shown via (a) Venn diagram, if the number of the analyzed studies are less than 3, and (b) UpSet Plot.

5. Finally, a volcano plot (Figure 13) showing the log-transformed p-values on the y-axis and the average log2-transformed fold change across all experiments on the x-axis is generated using `EnhancedVolcano` package. Again, the figure and the data used to create it are available for download.

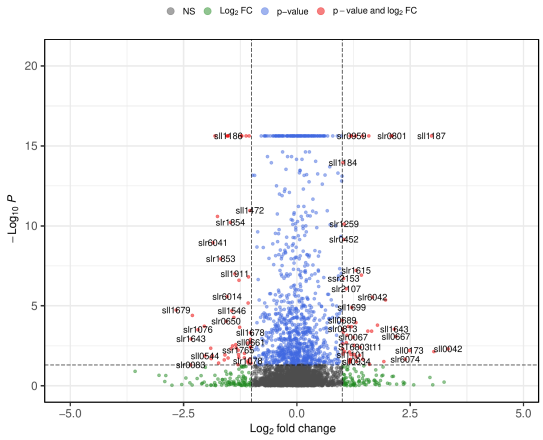
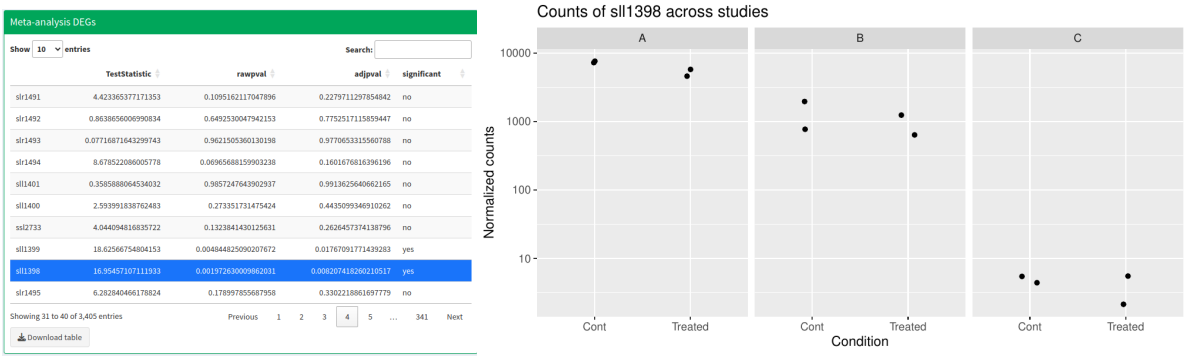


Figure 13: An example of the volcano plot generated by TMA.

6.3 DEGs Table & Count Plots

An interactive table (Figure 14a) of all the genes analyzed in the meta-analysis is created using the `DT` package. This table shows gene names, test statistic, raw p-values, and adjusted p-values, along with a column to token the significant genes with 'yes' for significant and 'no' for non-significant genes based on the the results of the meta-analysis. The table allows several sorting and search functions to facilitate data exploration and extraction. It is also could be downloaded as CSV file for further analysis. If the user would like to see the counts of any specified gene, highlighting the row of that gene will do the job and create the desired figure (Figure 14b) including counts of the highlighted gene under the two compared conditions across each individual study included in the analysis. This kind of plots might be helpful in examining the dosage effect on changes in expression levels of the gene.



(a) Table of DE analysis results

(b) Plots of Counts

Figure 14: The DEGs that was identified in the meta-analysis are shown in an interactive table (a) that allow users, for example, to search for specific genes or significant ones. Highlighting a row of the table will create a counts plot (b) of the normalized counts of the highlighted gene across all the studies included in the meta-analysis.

6.4 Functional Annotation

If the user decided to perform further functional annotation of the DEGs identified in the meta-analysis via GO and KEGG Pathway enrichment analyses, the app will generated two outputs for each one of the two analyses. Firstly, all the enriched GO terms or KEGG Pathways will be identified and be available to download as a CSV file. If the number of enriched GO terms or KEGG Pathways are less than 30, they will be shown in a lollipop figure that could be also downloaded. But if the number are more than 30, the top 30 (with the lowest FDR values) enriched GO terms or KEGG Pathways will be plotted only (Figure 15). The figures could also be downloaded as PDF files.

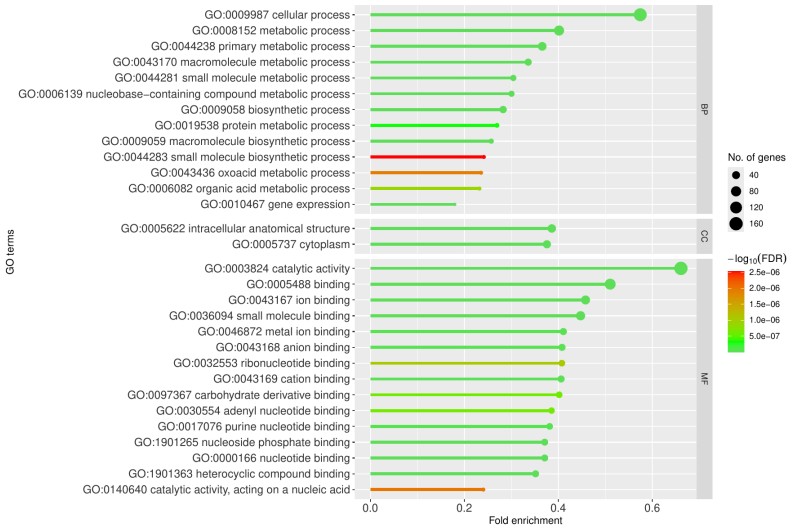


Figure 15: Enriched GO terms or KEGG Pathways are shown using lollipop figures. If the number of enriched terms or pathways are more than 30, the 30 terms with the lowest FDR values are drawn.

References

1. Abdel-Salam, E. M., Figueroa-Gonzalez, T. & Leister, D. Transcriptomic meta-analysis and functional validation identify genes linked to adaptation and involved in high-light acclimation in *Synechocystis* sp. PCC 6803. *Frontiers in Photobiology* **1**, (2023).
2. Paul, J. & Barari, M. Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? *Psychology & Marketing* **39**, 1099–1115 (2022).
3. Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* **15**, (2014).
4. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, (2014).
5. Fisher, R. A. Statistical methods for research workers. in *Breakthroughs in statistics: Methodology and distribution* (eds. Kotz, S. & Johnson, N. L.) 66–70 (Springer New York, New York, NY, 1992). doi:10.1007/978-1-4612-4380-9_6.
6. Mi, H. *et al.* Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols* **14**, 703–721 (2019).
7. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
8. Galili, T., O'Callaghan, A., Sidi, J. & Sievert, C. Heatmaply: An r package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* **34**, 1600–1602 (2017).