Journal of Visualized Experiments

Analyzing multifactorial RNA-Seq experiments with DicoExpress

--Manuscript Draft--

Article Type:	Invited Methods Collection - JoVE Produced Video
Manuscript Number:	JoVE62566R2
Full Title:	Analyzing multifactorial RNA-Seq experiments with DicoExpress
Corresponding Author:	Marie-Laure Martin-Magniette IPS2: Institut des Sciences des Plantes de Paris Saclay Gif-sur-Yvette, FRANCE
Corresponding Author's Institution:	IPS2: Institut des Sciences des Plantes de Paris Saclay
Corresponding Author E-Mail:	marie_laure.martin-magniette@agroparistech.fr
Order of Authors:	Kevin Baudry
	Christine Paysant-Le Roux
	Stefano Colella
	Benoît Castandet
	Marie-Laure Martin-Magniette
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please specify the section of the submitted manuscript.	Engineering
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Gif-sur-Yvette, France
Please confirm that you have read and agree to the terms and conditions of the author license agreement that applies below:	I agree to the Author License Agreement
Please provide any comments to the journal here.	
Please indicate whether this article will be Standard Access or Open Access.	Standard Access ($1400)
Please confirm that you have read and agree to the terms and conditions of the video release that applies below:	I agree to the Video Release

JOURNAL OF VISUALIZED EXPERIMENTS
1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

Standard Manuscript Template

TITLE:

Analyzing multifactorial RNA-Seq experiments with DicoExpress

AUTHORS AND AFFILIATIONS:

Kevin Baudry[1,2,3], Christine Paysant- Le Roux[1,2], Stefano Colella[4], Benoît Castandet[1,2], Marie-Laure Martin-Magniette[1,2,5]

1. Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Orsay, France.
2. Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris Saclay (IPS2), Orsay, France.
3. Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Gif-sur-Yvette, France.
4. LSTM, Univ Montpellier, INRAE, IRD, CIRAD, Institut Agro, Montpellier, France.
5. Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, Paris, France.

Correspondence to: Marie-Laure Martin-Magniette at marie-laure.magniette@inrae.fr

SUMMARY:

DiCoExpress is a script-based tool implemented in R to perform an RNA-Seq analysis from quality control to co-expression. DiCoExpress handles complete and unbalanced design up to 2 biological factors. This video tutorial guides the user through the different features of DiCoExpress.

ABSTRACT:

The proper use of statistical modeling in NGS data analysis requires an advanced level of expertise. There has recently been a growing consensus on using generalized linear models for differential analysis of RNA-Seq data and the advantage of mixture models to perform co-expression analysis. To offer a managed setting to use these modeling approaches, we developed DiCoExpress that provides a standardized R pipeline to perform an RNA-Seq analysis. Without any particular knowledge in statistics or R programming, beginners can perform a complete RNA-Seq analysis from quality controls to co-expression through differential analysis based on contrasts inside generalized linear models. An enrichment analysis is proposed both on the lists of differentially expressed genes, and the co-expressed gene clusters. This video tutorial is conceived as a step-by-step protocol to help users take full advantage of DiCoExpress and its potential in empowering the biological interpretation of an RNA-Seq experiment.

INTRODUCTION:

Next-generation RNA sequencing (RNA-Seq) technology is now the gold standard of transcriptome analysis[1]. Since the early days of the technology, the combined efforts of bioinformaticians and biostatisticians have resulted in the development of numerous methods tackling all the essential steps of transcriptomic analyses, from mapping to transcript quantification[2]. Most of the tools available today to the biologist are developed within the R software environment for statistical computing and graphs[3], and many packages for biological

data analysis are available in the Bioconductor repository[4]. These packages offer total control and customization of the analysis, but they come at the cost of extensive use of a command-line interface. Because many biologists are more comfortable with a "point and click" approach[5], the democratization of RNA-Seq analyses requires the development of more user-friendly interfaces or protocols[6]. For example, it is possible to build web interfaces of R packages using Shiny[7], and command-line data analysis is made more intuitive with the R-studio[8] interface. The development of dedicated, step-by-step tutorials can also help the novel user. In particular, a video tutorial supplements a classic text one, leading to a deeper understanding of all the procedure steps.

We recently developed DiCoExpress[9], a tool for analyzing multifactorial RNA-Seq experiments in R using methods considered to be the best ones based on neutral comparison studies[10–12]. Starting from a count table, DiCoExpress proposes a data quality control step followed by generalized linear models (GLM) differential gene expression analysis (edgeR package[13]) and the generation of co-expression clusters using Gaussian mixture models (coseq package[12]). DiCoExpress handles complete and unbalanced design up to 2 biological factors (i.e., genotype and treatment) and one technical factor (i.e., replicate). The originality of DiCoExpress lies in its directory architecture storing and organizing data, scripts, and results and in the automation of the writing of the contrasts allowing the user to investigate numerous questions within the same statistical model. An effort was also made to provide graphical outputs illustrating the statistical results.

The DiCoExpress workspace is available at https://forgemia.inra.fr/GNet/dicoexpress. It contains four directories, two pdf, and two text files. The Data/ directory contains the input datasets; for this protocol, we will use the "tutorial" dataset. The Sources/ directory contains seven R functions necessary to perform the analysis, and must not be modified by the user. The analysis is run using scripts stored in the Template_scripts/ directory. The one used in this protocol is called DiCoExpress_Tutorial_JoVE.R and can be easily adapted to any transcriptomic project. All the results are written in the Results/ directory and stored in a subdirectory named according to the project. The README.md file contains useful installation information, and any specific details concerning the method and its use can be found in the DiCoExpress_Reference_Manual.pdf file.

This video tutorial guides the user through the different features of DiCoExpress with the aim to overcome the reluctance felt by biologists using command-line-based tools. We present here the analysis of an artificial RNA-Seq dataset describing gene expression in three biological replicates of four genotypes, with or without treatment. We will now go through the different steps of the DiCoExpress workflow illustrated in Figure 1. The script described in the Protocol section and input files are available on the project site:
https://forgemia.inra.fr/GNet/dicoexpress

PREREQUISITES

Prepare data files

The four csv files stored in the Data/ directory should be named according to the project name. In our example, all the names, therefore, begin with "Tutorial", and we will set `Project_Name =` "Tutorial" in Step 4 of the protocol. The separator used in the csv files must be indicated in the Sep variable in Step 4. In our "tutorial" dataset, the separator is a tabulation. For advanced users the full dataset can be reduced to a subset by providing a list of instructions and a new Project_Name through the Filter variable. This option avoids redundant copies of the input files and verifies FAIR principles[14].

Among the four csv files, only the COUNTS and TARGET files are mandatory. They contain the raw counts for every gene (here Tutorial_COUNTS.csv) and the experimental design description (here Tutorial_TARGET.csv). The TARGET.csv file describes every sample (one sample per row) with a modality for each biological or technical factor (in the columns). We strongly recommend that the names chosen for the modalities start with a letter, not a number. The name of the last column ("Replicate") cannot be changed. Finally, the sample names (first column) must match the names in the headings of the COUNTS.csv file (Genotype1_control_rep1 in our example). The Enrichment.csv file in which every line contains one Gene_ID and one annotation term is only required if the user plans to run the enrichment analysis. If one gene has several annotations, they will have to be written on different lines. The Annotation.csv file is optional and is used to add a short description of every gene in the output files. The best way to get an annotation file is to retrieve the information from dedicated databases (e.g., Thalemine: https://bar.utoronto.ca/thalemine/begin.do for Arabidopsis).

Installation of DiCoExpress

DiCoExpress requires specific R packages. Use the command line source("../Sources/Install_Packages.R") in the R console to check the required packages' installation status. For users on Linux, another solution is to install the container dedicated to DiCoExpress and available at https://forgemia.inra.fr/GNet/dicoexpress/container_registry. By definition, this container contains DiCoExpress with all of the parts needed, such as libraries and other dependencies.

PROTOCOL:

1. **DiCoExpress**

1.1. Open a R studio session and set directory to Template_scripts.

1.2. Open the DiCoExpress_Tutorial.R script in R studio .

1.3. Load DiCoExpress functions in the R session with the following commands:

```
> source("../Sources/Load_Functions.R")
> Load_Functions()
> Data_Directory = "../Data"
> Results_Directory = "../Results/"
```

1.4. Load data files in the R session with the following commands:

```
> Project_Name = "Tutorial"
> Filter = NULL
> Sep="\t"
> Data_Files = Load_Data_Files(Data_Directory,
Project_Name, Filter, Sep)
```

1.5. Split the object Data_Files in several objects to manipulate them easily:

```
> Project_Name = Data_Files$Project_Name
> Target = Data_Files$Target
> Raw_Counts = Data_Files$Raw_Counts
> Annotation = Data_Files$Annotation
> Reference_Enrichment = Data_Files$Reference_Enrichment
```

1.6. Choose a strategy among "NbConditions", "NbReplicates" or "filterByExpr" and a threshold to filter low expressed genes. Here we choose

```
> Filter_Strategy = "NbReplicates"
> CPM_Cutoff = 1
```

1.7. Specify group colors with the command

```
>Color_Group = NULL
```

NOTE: When it is set to NULL, R automatically attributes colors to the biological conditions. Otherwise enter a vector indicating a color per biological group.

1.8. Choose a normalization method among those accepted by the function calcNormFactors of edgeR. As for example

```
> Normalization_Method = "TMM"
```

1.9. Perform the quality control by executing the following function

```
> Quality_Control(Data_Directory, Results_Directory,
Project_Name, Target, Raw_Counts, Filter_Strategy, Color_Group,
CPM_Cutoff, Normalization_Method)
```

1.10. State `Replicate = TRUE` if data are paired according to the replicate factor, `FALSE` otherwise.

1.11. Assign `Interaction = TRUE` to consider an interaction between the two biological factors, `FALSE` otherwise.

1.12. Specify the statistical model with the following commands

```
> Model = GLM_Contrasts(Results_Directory, Project_Name,
Target, Replicate, Interaction)
> GLM_Model = Model$GLM_Model
> Contrasts = Model$Contrasts
```

1.13. Define the threshold of the False Discovery Rate, here 0.05

```
> Alpha_DiffAnalysis =0.05
```

1.14. Perform the differential analysis with the following commands

```
> Index_Contrast=1:nrow(Contrasts)
> NbGenes_Profiles = 20
> NbGenes_Clustering = 50
> DiffAnalysis.edgeR(Data_Directory, Results_Directory,
Project_Name, Target, Raw_Counts, GLM_Model, Contrasts,
Index_Contrast, Filter_Strategy, Alpha_DiffAnalysis,
NbGenes_Profiles, NbGenes_Clustering, CPM_Cutoff,
Normalization_Method)
```

1.15. Fix a threshold for the enrichment analysis, here 0.01

```
> Alpha_Enrichment = 0.01
```

1.16. Perform the enrichment analysis of differentially expressed genes (DEG) lists

```
> Title = NULL
> Enrichment(Results_Directory, Project_Name, Title,
Reference_Enrichment, Alpha_Enrichment)
```

1.17. Choose DEG lists to be compared. As for example,

```
> Groups = Contrasts$Contrasts[24:28]
```

1.18. Provide a name for the list comparison. This name is used for the directory where the output files will be saved

```
> Title = "Interaction_with_Genotypes_1_and_2"
```

1.19. Specify the action to be done on the DEG lists by setting the parameter Operation to union or intersection. We choose

```
> Operation = “Union”
```

1.20. Compare the DEGs lists

```
> Venn_IntersectUnion(Data_Directory, Results_Directory,
Project_Name, Title, Groups, Operation)
```

1.21. Perform a co-expression analysis with the function

```
> Coexpression_coseq(Data_Directory, Results_Directory,
Project_Name, Title, Target, Raw_Counts, Color_Group)
```

1.22. Perform enrichment analysis of the co-expression clusters

```
> Enrichment(Results_Directory, Project_Name, Title,
Reference_Enrichment, Alpha_Enrichment)
```

1.23. Generate two log files containing all the necessary information to reproduce the analysis

```
> Save_Parameters()
```

NOTE: Command lines used in this protocol are shown in **Figure 2**. Lines that have to be modified to analyze another dataset are highlighted.

REPRESENTATIVE RESULTS:

All the DiCoExpress outputs are saved in the *Tutorial/* directory, itself placed within the *Results/* directory. We provide here some guidance for assessing the overall quality of the analysis.

Quality Control

The quality control output, located in the *Quality_Control/* directory, is essential to verify that the RNA-Seq analysis results are reliable. The Data_Quality_Control.pdf file contains several plots obtained with raw and normalized data that can be used to identify any potential issues with the data. The total normalized counts per sample should be similar when comparing both intra- and inter-conditions. Moreover, the normalized gene expression counts are expected to exhibit similar median and variance both in intra- and inter-conditions (**Figure 3A**). Otherwise, this could be the sign of non-similar variance between conditions, an issue that could be problematic for model fitting.

Finally, PCA plots on normalized counts produced in DiCoExpress are helpful to identify potential underlying data structures (**Figure 3B**). In our example, there is no clustering according to the replicates, meaning that this factor is not discriminant. At the same time, a clear distinction can be identified between treatments. These results indicate a good quality dataset since the biological effect is always expected to be stronger than the replicate one. In conclusion, the overall quality observed here does not prevent any subsequent analysis of the entire dataset.

Statistical modeling

DiCoExpress facilitates the writing of the statistical modeling of the logarithm of the mean expression from the two variables *Replicate* and *Interaction*. A replicate effect is conceivable if the samples of all the biological conditions are collected at the same time and that this experiment is replicated on different days to measure biological variability. In a typical plant science experiment, for example, samples are grown in the same growth chamber regardless of the biological condition under study and biological replicates correspond to experiments started at different days. In this case, the samples of the same replicate are paired, and you should set *Replicate* to TRUE. Otherwise, *Replicate* should be set to FALSE. This replicate effect is also known as a batch effect.

If the experimental design is described by two biological factors expected to interact, set the variable `Interaction` to TRUE to consider the interaction. Note that for a project containing only one biological factor, the variable `Interaction` is automatically set to FALSE.

Differential Analysis

The DEG identified for all the tested contrasts are available in text files located in their respective subdirectories within the *DiffAnalysis/* directory. By default, all the contrasts are tested. Depending on the experimental design, some contrasts can be of limited biological interest (for example, an average on several genotypes). Note that the false positive control is performed per contrast ensuring that potentially irrelevant contrasts do not impact the analysis. It is however possible to produce plots only containing the contrast of interests by acting on the *Index_Contrast* variable. Details are available in the online reference manual.
It is essential to notice that *DiffAnalysis/* also contains the raw p-value histograms that have recently been shown to be the best way to assess the quality of the modeling [11]. The expected distribution of raw p-values is supposed to be uniform, with possibly a peak at the left end side of the distribution. A high peak for a raw p-value of 1 is indicative of model fitting issues. In this case, the problem can often be solved by increasing the set `CPM_Cutoff` value, for example, from 1 to 5. Examples of raw histograms are available in **Figure 4A** and in https://forgemia.inra.fr/GNet/dicoexpress/-/blob/master/DiCoExpress_Tutorial.pdf.
For every tested contrast, expression profiles of the top DEG identified (top 20 by default) are plotted in the file Top20_Profile.pdf located in the directory of the contrast. An example for one gene identified as DE in one contrast is shown in **Figure 4B**. The number of up and down DEG is plotted for every tested contrast and is found in the file Down_Up_DEG.pdf (example in Figu**re 4C**).

Co-expression Analysis

In our example, the co-expression analysis is performed on the union of 5 DEG lists, identified by contrast looking for treatment response variation between genotype 1 or 2 against others. Venn diagram of DEG is shown in **Figure 5A**. The co-expressed genes for every identified cluster are printed in individual text files (one file per cluster). The expression profiles of the different clusters together are available in the Boxplot_profiles_Coseq.pdf file (see example in **Figure 5B**). Although customization options are available, they should only be used by advanced users. Please refer to the reference manual for a complete explanation of the different parameters.

Enrichment Analysis

Lists corresponding to the contrast and cluster enrichment analyses are located in their respective directories. An annotation term found as significant in this analysis can be either over- or under-represented in the Gene_ID list. This information is included in the output file. Note that the test decision is made from the raw p-values. If the user wants to adjust the raw p-values a posteriori, they are available in the file with suffix All_Enrichment_Results.txt.

Validity of DiCoExpress

Although DiCoExpress has been developed to facilitate multifactorial RNA-Seq experiments analyses, the validity of its results largely depends on the characteristics of the dataset. Several outputs should carefully be checked before any valid interpretation of the results. First, in the quality control step, the normalized library size should be similar and the normalized gene expression count should exhibit similar median and variance in both intra- and inter-conditions. Then, a particular attention should be paid to the shape of the raw p-value histograms. Finally,

when performing a co-expression analysis, a clearly defined minimum value for the ICL is indicative of a good quality. It these conditions are not met, any interpretation of the results is likely to be erroneous.

FIGURE AND TABLE LEGENDS:

Figure 1. **The DiCoExpress Analysis pipeline.**
The seven steps of a complete RNA-Seq analysis using DiCoExpress are indicated blue boxes represent steps where statistical methods are performed. Step 7 (Enrichment) can be done after Step 4 (Differential Analysis and is named 7.1 in Figure 2) and/or Step 6 (Co-expression analysis and is named 7.2 in Figure 2). Red numbers correspond to the step numbers in the protocol.

Figure 2. Screenshots of DiCoExpress command lines.
Command lines used to analyze the tutorial dataset are indicated. Number in black circles are the same than in figure 1. Red rectangles highlight lines that can be customized by the user.

Figure 3: Representative results of the quality control step.
Figure obtained with the "Tutorial" dataset normalized counts. **A**) Boxplot of normalized counts. **B**) PCA on normalized counts.

Figure 4: Representative results of the Differential expression analysis
Figure obtained with the "Tutorial" dataset. **A)** Raw p-value histogram of the [control_Genotype2 – control_Genotype3] contrast. **B)** C1G62301.1 gene expression profile in every genotype and condition, one of the Top20 Differentially Expressed Gene in the [control_Genotype2 – control_Genotype3] contrast. **C)** Number of up and down Differentially Expressed Genes in every tested contrast.

Figure 5: Representative results of the Coexpression Analysis
Figure obtained with the "Tutorial" dataset. **A**) Venn diagram of DEG from the 5 "interaction with Genotype 1 and 2" contrasts. DEG from the treatment response variation between Genotype 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4 are in circle A, B, C, D, E respectively. The number written at the bottom right ("14877") is the number of genes that are not DE in any list. **B**) Expression Profile of genes from the coexpression Cluster 3. Figure is extracted from Tutorial_Interaction_with_Genotypes_1_and_2_Boxplot_profiles_Coseq.pdf

DISCUSSION:
Because RNA-Seq has become a ubiquitous method in biological studies, there is a constant need to develop versatile and user-friendly analytical tools. A critical step within most of the analytical workflows is often to identify with confidence the genes differentially expressed between biological conditions and/or treatments[15]. The production of reliable results requires proper statistical modeling, which has been the motivation for the development of DiCoExpress.

DiCoExpress is a script-based tool implemented in R that aims at helping biologists take full advantage of the possibilities of neutral comparison studies when looking for DEG. DiCoExpress provides a standardized pipeline offering the opportunity to evaluate the data structure and quality, therefore ensuring the best modeling approach is chosen. Without any particular knowledge in statistics or R programming, it allows beginners to perform a complete RNA-Seq analysis from quality controls to co-expression through differential analysis based on contrasts inside generalized linear models. It is important to note that DiCoExpress focuses on the statistical part of an RNA-Seq analysis and requires a count table as input. The multiple bioinformatics methods dedicated to RNA-Seq read alignments and the creation of count tables are out of the scope of the tool. They nevertheless have a direct influence on the quality of the final analysis and should be carefully chosen.

Although DiCoExpress is not a "point and click" tool, its directory architecture and the template script provided and used in the R-Studio interface make it accessible to biologists with minimal knowledge of R. Once DiCoExpress is installed, users should know how to use a function in R and identify required and optional arguments. The first critical step is to correctly provide the two mandatory files containing the raw counts for every gene (the COUNTS file) and the experimental design description (the TARGET file). The used separator should be the same for every file and the description of the samples should be done appropriately according to the modalities of the biological factors. Once the two files are loaded in DiCoExpress, the analysis is almost automated until the second critical step, i.e., the co-expression analysis. This analysis can indeed be time-consuming and a powerful calculation server could be required to run it on large datasets.

Because automation of the contrast writing becomes challenging for more than two biological factors, we limited DiCoExpress to the complete and unbalanced design of up to 2 biological factors. If a project contains more than 2 biological factors, a practical solution is to collapse two of the initial factors to create a new one. Nevertheless, one has to keep in mind that the difficulty of giving a meaningful biological interpretation increases when the biological factor number increases.

DiCoExpress is conceived as an evolving tool and we strongly encourage users to subscribe to the mailing list (https://groupes.renater.fr/sympa/subscribe/dicoexpress). Any modifications or improvements to the tool will be announced on the list and we welcome questions or suggestions. We also hope that the adoption of DiCoExpress by a large community will allow tracking and fixing any bugs that might occur in some particular analysis context. All the updates and corrections will be pushed to the git directory https://forgemia.inra.fr/GNet/dicoexpress.

ACKNOWLEDGMENTS:
This work was mainly supported by the ANR PSYCHE (ANR-16-CE20-0009). The authors thank F. Desprez for the construction of the container of DiCoExpress. KB work is supported by the Investment for the Future ANR-10-BTBR-01-01 Amaizing program. The GQE and IPS2 laboratories benefit from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007).

DISCLOSURES:

The authors have nothing to disclose

REFERENCES:

1. Wang, Z., Gerstein, M., Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*. **10** (1), 57–63, doi: 10.1038/nrg2484 (2009).

2. Yang, I.S., Kim, S. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics*. **13** (4), 119–125, doi: 10.5808/gi.2015.13.4.119 (2015).

3. R Core Team (2020) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* at <https://www.R-project.org/> (2020).

4. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*. **12** (2), 115–121, doi: 10.1038/nmeth.3252 (2015).

5. Smith, D.R. The battle for user-friendly bioinformatics. *Frontiers in Genetics*. **4**, 187, doi: 10.3389/fgene.2013.00187 (2013).

6. Pavelin, K., Cham, J.A., Matos, P. de, Brooksbank, C., Cameron, G., Steinbeck, C. Bioinformatics Meets User-Centred Design: A Perspective. *PLoS Computational Biology*. **8** (7), e1002554, doi: 10.1371/journal.pcbi.1002554 (2012).

7. Shiny: web application framework. at <https ://rdrr.io/cran/shiny/> (n.d.).

8. RStudio Team (2020). RStudio: Integrated Development for R. *RStudio, PBC, Boston, MA*. at <http://www.rstudio.com/> (n.d.).

9. Lambert, I., Roux, C.P.-L., Colella, S., Martin-Magniette, M.-L. DiCoExpress: a tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. *Plant methods*. **16** (1), 68, doi: 10.1186/s13007-020-00611-7 (2020).

10. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*. **14** (6), 671–83, doi: 10.1093/bib/bbs046 (2012).

11. Rigaill, G. *et al.* Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in Bioinformatics*. **19** (1), bbw092, doi: 10.1093/bib/bbw092 (2016).

12. Rau, A., Maugis-Rabusseau, C. Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*. **19** (3), bbw128, doi: 10.1093/bib/bbw128 (2017).

13. Robinson, M.D., McCarthy, D.J., Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26** (1), 139–140, doi: 10.1093/bioinformatics/btp616 (2009).

14. Wilkinson, M.D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. **3** (1), 160018, doi: 10.1038/sdata.2016.18 (2016).

15. Stark, R., Grzelak, M., Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics*. **20** (11), 631–656, doi: 10.1038/s41576-019-0150-2 (2019).

Figure1

Click here to access/download;Figure;DiCoExpress_Figure_1.pdf

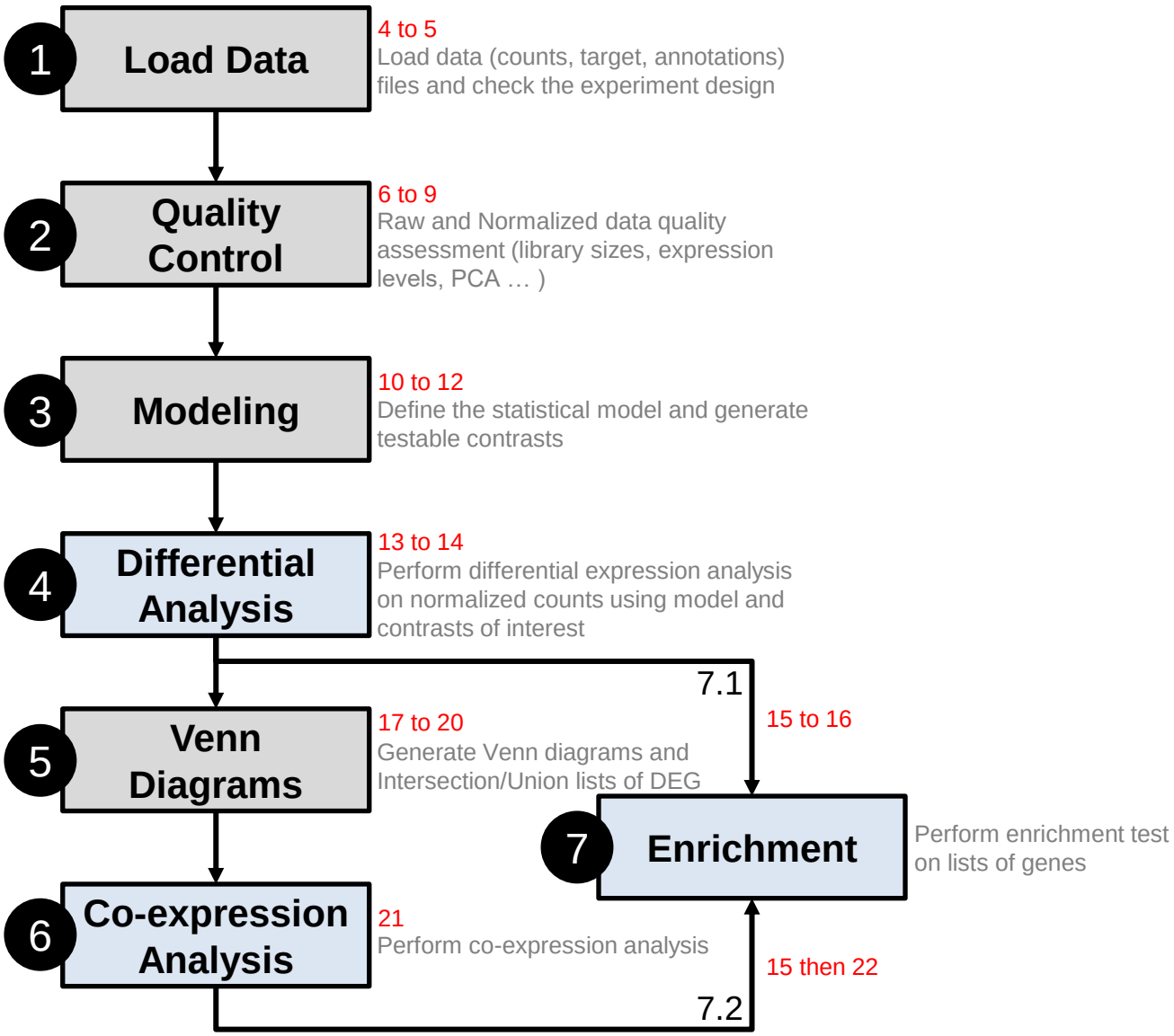


Figure2

Click here to access/download;Figure;DiCoExpress_Figure_2.pdf

```
> ## To begin ####
> source("../Sources/Load_Functions.R")
> Load_Functions()
>
> Data_Directory = "../Data"
> Results_Directory = "../Results/"
>
> ## Data description and importation
> Project_Name = "Tutorial"
> Filter = NULL
> Sep = "\t"
>
> Data_Files = Load_Data_Files(Data_Directory, Project_Name, Filter, Sep)
>
> Project_Name = Data_Files$Project_Name
> Target = Data_Files$Target
> Raw_Counts = Data_Files$Raw_Counts
> Annotation = Data_Files$Annotation
> Reference_Enrichment = Data_Files$Reference_Enrichment
>
> ## Quality controls
> Filter_Strategy = "NbReplicates"
> CPM_Cutoff = 1
> Color_Group = NULL
> Normalization_Method = "TMM"
>
> Quality_Control(Data_Directory, Results_Directory, Project_Name, Target,
+                 Raw_Counts, Filter_Strategy, Color_Group, CPM_Cutoff,
+                 Normalization_Method)
>
> ##Statistical model
> Replicate = TRUE
> Interaction = TRUE
>
> Model = GLM_Contrasts(Results_Directory, Project_Name, Target, Replicate, Interaction)
> GLM_Model = Model$GLM_Model
> Contrasts = Model$Contrasts
>
> ## Differential analysis
> Alpha_DiffAnalysis = 0.05
> Index_Contrast = 1:nrow(Contrasts)
> NbGenes_Profiles = 20
> NbGenes_Clustering = 50
>
> DiffAnalysis.edgeR(Data_Directory, Results_Directory, Project_Name,
+                    Target, Raw_Counts, GLM_Model, Contrasts, Index_Contrast,
+                    Filter_Strategy, Alpha_DiffAnalysis, NbGenes_Profiles,
+                    NbGenes_Clustering,
+                    CPM_Cutoff, Normalization_Method)
>
> ## DEG enrichment
> Alpha_Enrichment = 0.01
> Title=NULL
> Enrichment(Results_Directory, Project_Name, Title, Reference_Enrichment, Alpha_Enrichment)
>
> ## Venn diagrams
> Title="Interaction_with_Genotypes_1_and_2"
> Groups = Contrasts$Contrasts[24:28]
> Operation="Union"
>
> Venn_IntersectUnion(Data_Directory, Results_Directory, Project_Name, Title,
+                     Groups, Operation)
>
> ## Co-expression
> Coexpression_coseq(Data_Directory, Results_Directory, Project_Name,
+ Title, Target, Raw_Counts, Color_Group)
>
> ## Cluster enrichment
> Enrichment(Results_Directory, Project_Name, Title, Reference_Enrichment, Alpha_Enrichment)
>
> ## Save parameters and package versions
> Save_Parameters()
```

1
2
3
4
7.1
5
6
7.2

Figure3

Click here to access/download;Figure;DiCoExpress_Figure_3.pdf

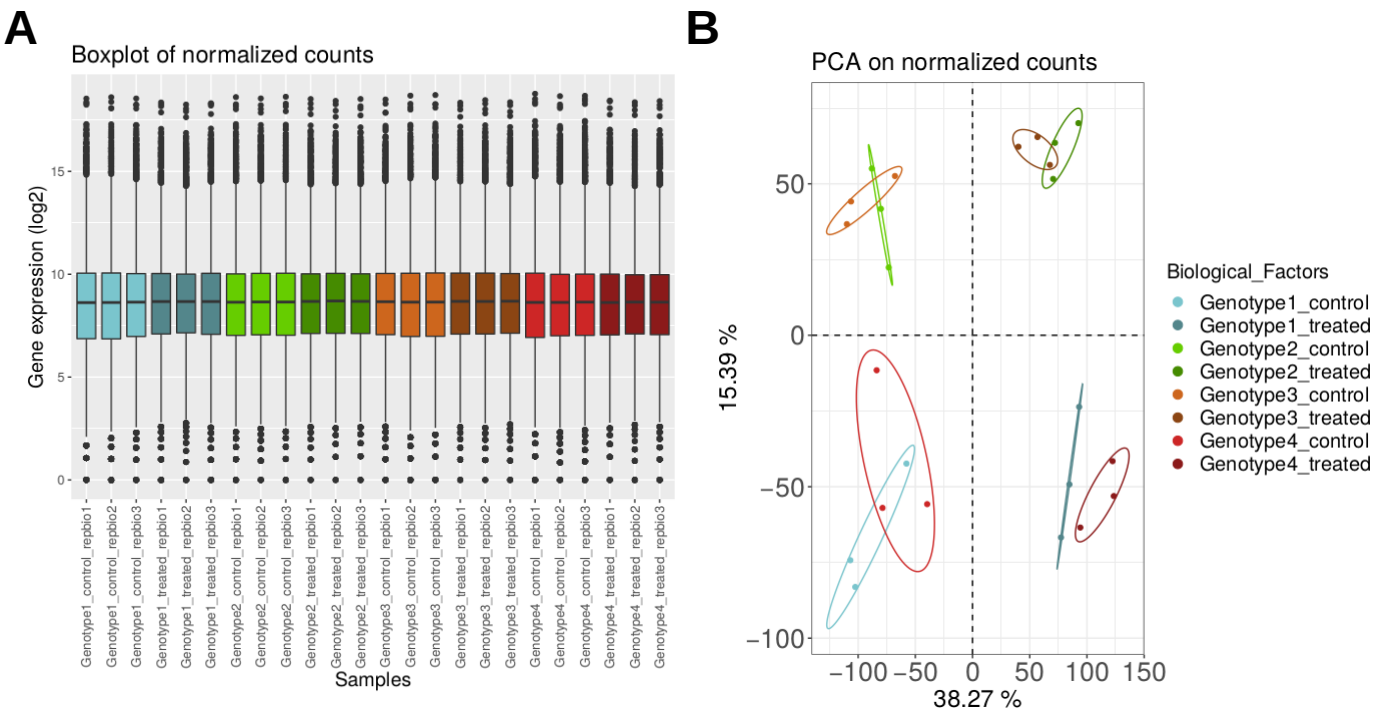


Figure4

Click here to access/download;Figure;DiCoExpress_Figure_4.pdf

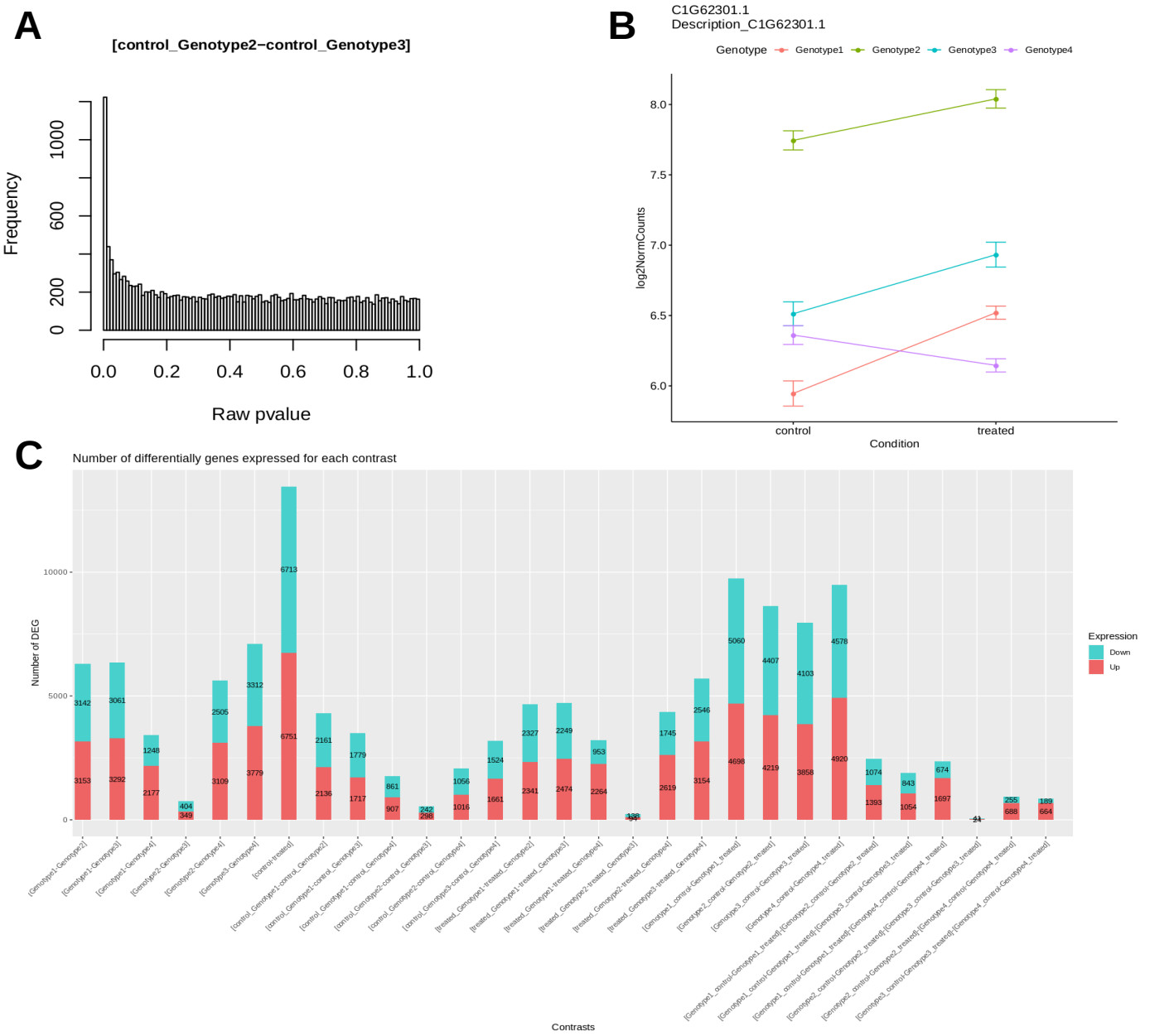
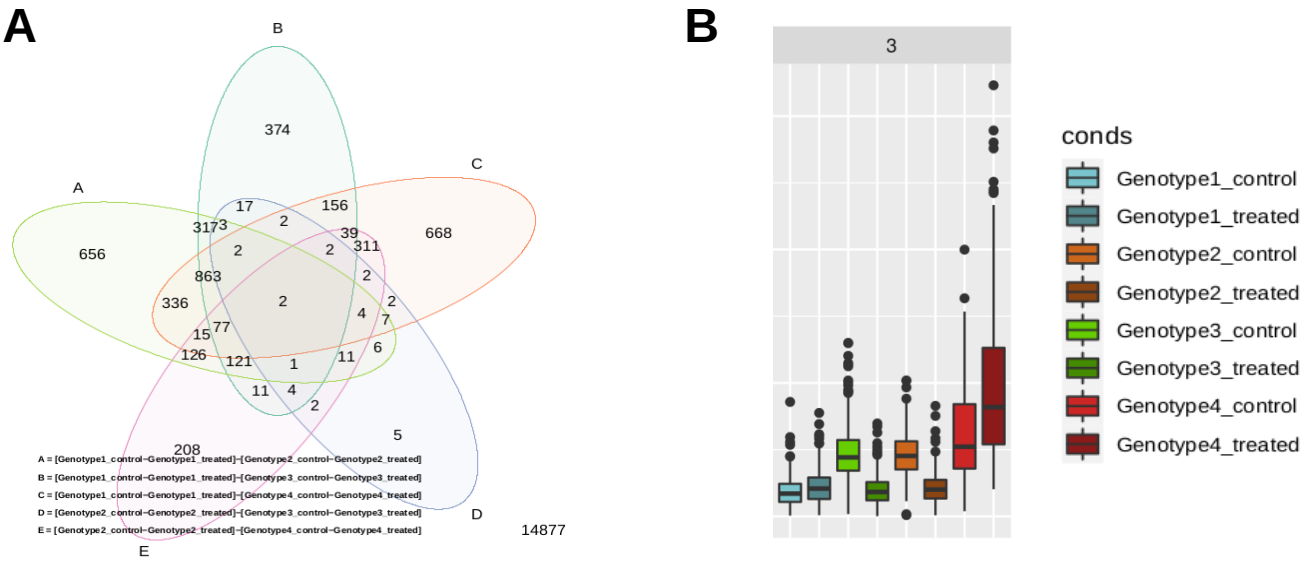


Figure5

Click here to access/download;Figure;DiCoExpress_Figure_5.pdf



IPS2
Institute of Plant Sciences
Paris - Saclay

université
PARIS-SACLAY

Paris, 21th June 2021

Dear Editor,

We thank you for your feedback and for extending the deadline for the submission of our revised manuscript, now entitled: **"Analyzing multifactorial RNA-Seq experiments with DicoExpress"**
We wish to thank the reviewers for their comments and new suggestions. We modified the manuscript following your and their recommendations. We proofread the manuscript, included 3 new figures and added a paragraph about the validity of DiCoExpress.

We hope that the manuscript in the present revised form will meet the standards for publication in JoVE.

Concerning the video, we would be grateful if you could give us more explanations of what you expect to get. As proposed by one reviewer, we will include a step to explain how to install DiCoExpress before a first use.

Best regards,

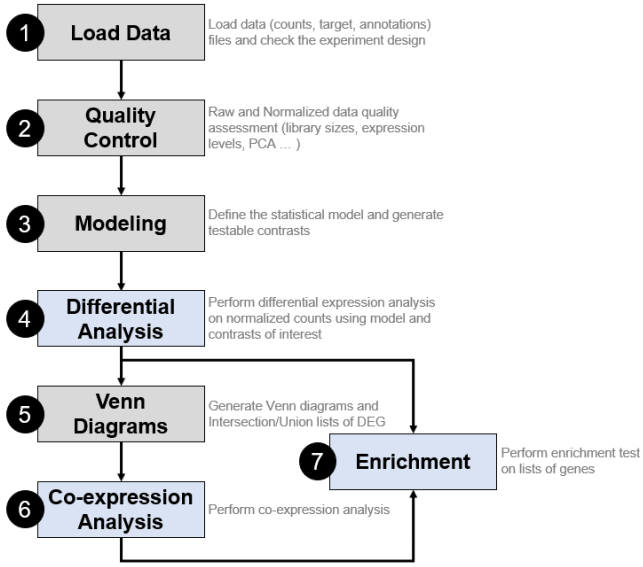
Marie-Laure Martin-Magniette

Institute of Plant Sciences Paris-Saclay (IPS2)
Recherche Fondamentale – Enseignement Supérieur – Innovation en sciences végétales
Plateau du Moulon - Rue Noetzlin Bât 630 CS 80004 – 91192 GIF SUR YVETTE Cedex
Tél. 01.69.15.33.30 – Fax : 09 72 56 17 74

Proposal for the sketch de la video

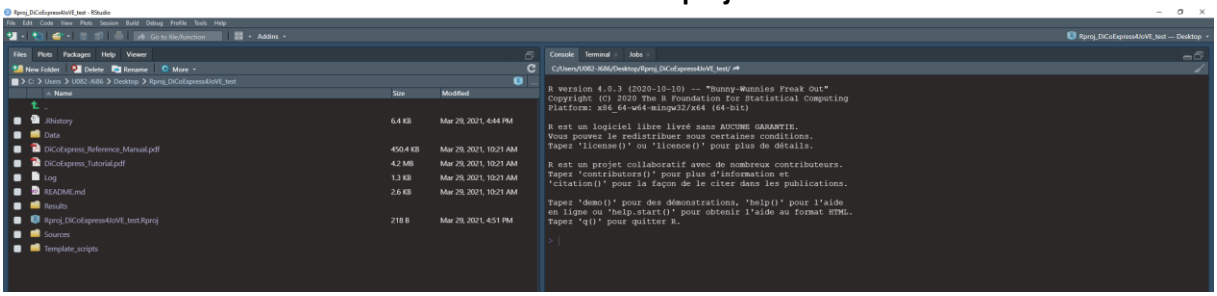
INTRODUCTION

General presentation of DiCoExpress and its advantages by using the workflow.
Description of the public and the skills required to use DiCoExpress

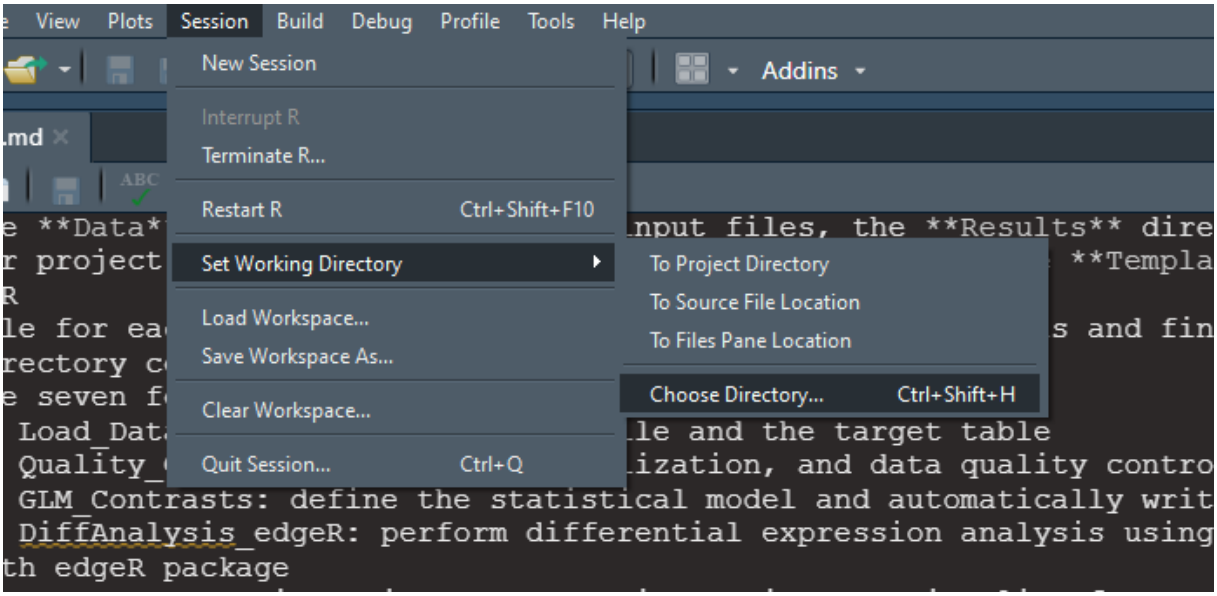


GETTING STARTED WITHDICOEXPRESS IN R-STUDIO

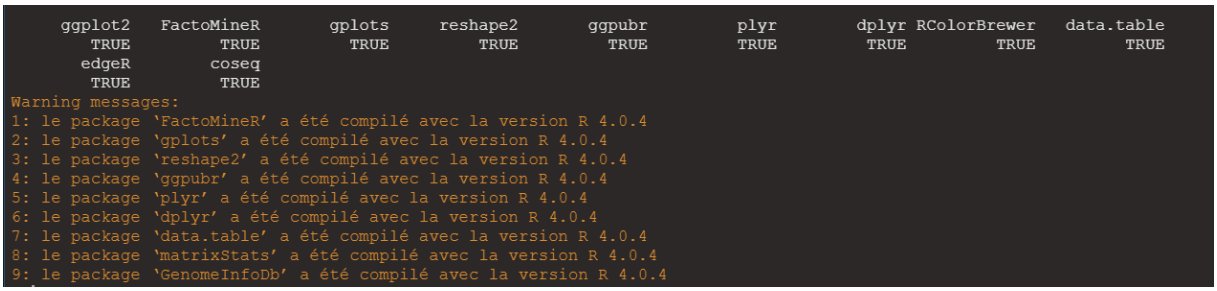
We do the video tutorial in R-Studio as it will allow in the same software to show the directory structure and also launch the commands without typing everyrhing.
The video will be a voice over **screen of R-Studio in an R-project**:



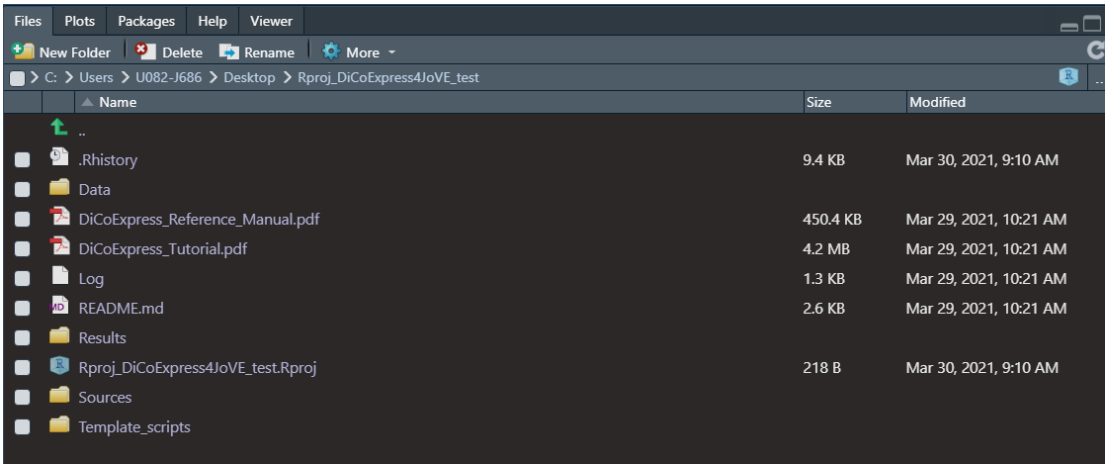
1. First we'll need to set the working directory to Template script



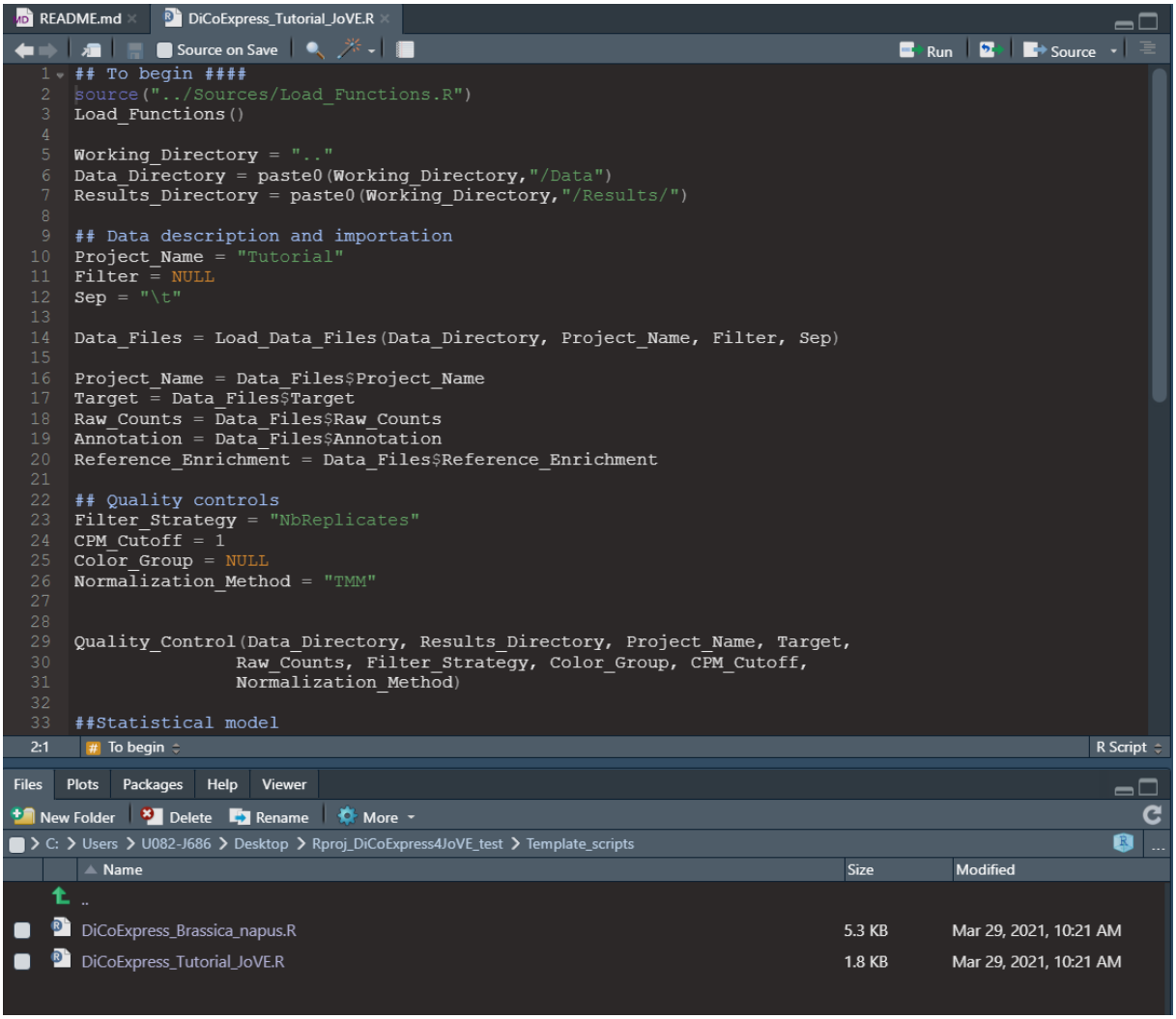
2. **We'll do package installation tests**



3. **Showing the directory structure in the folder**

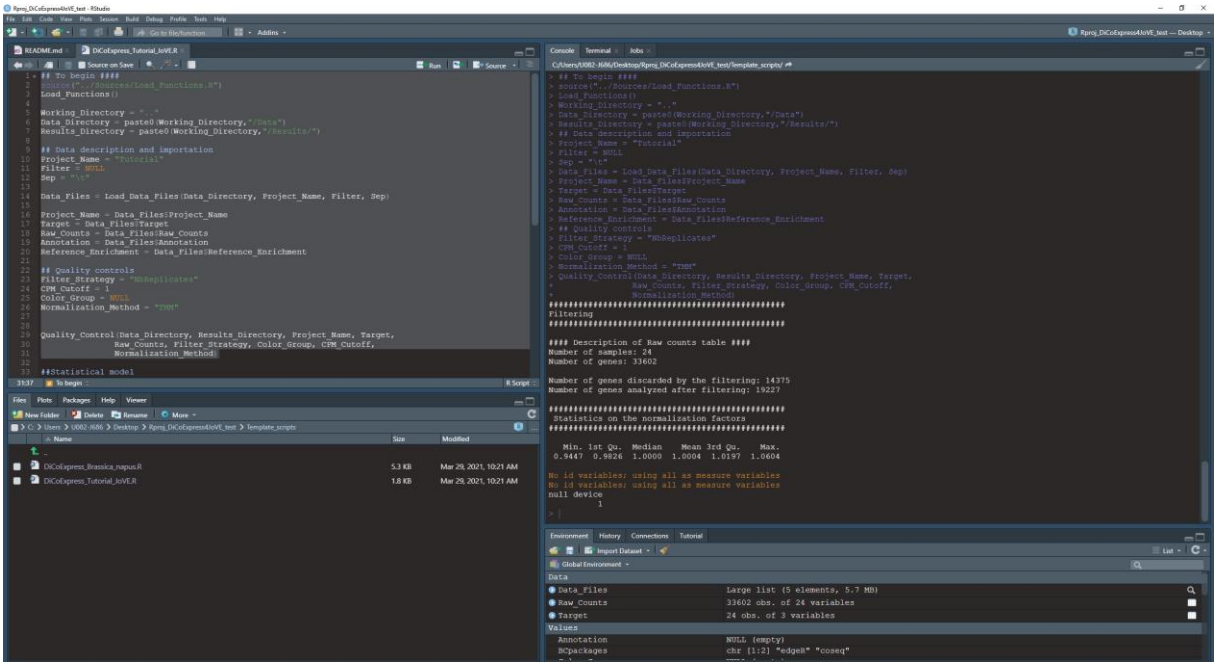


4. **Open "Template script"**

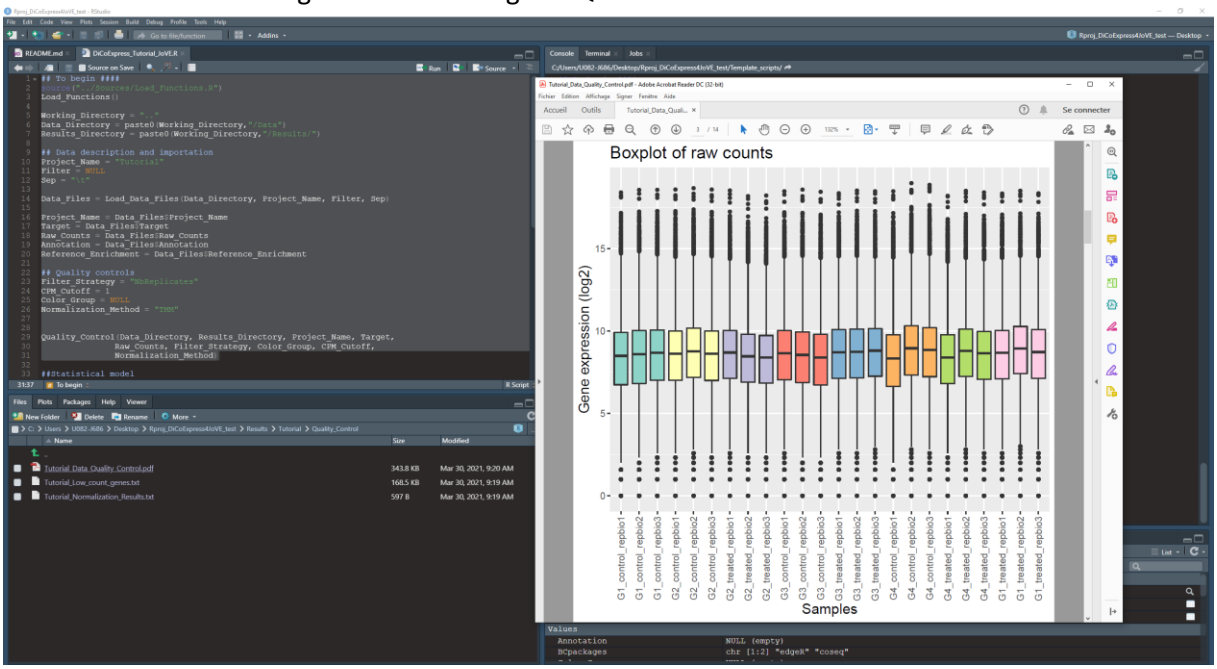


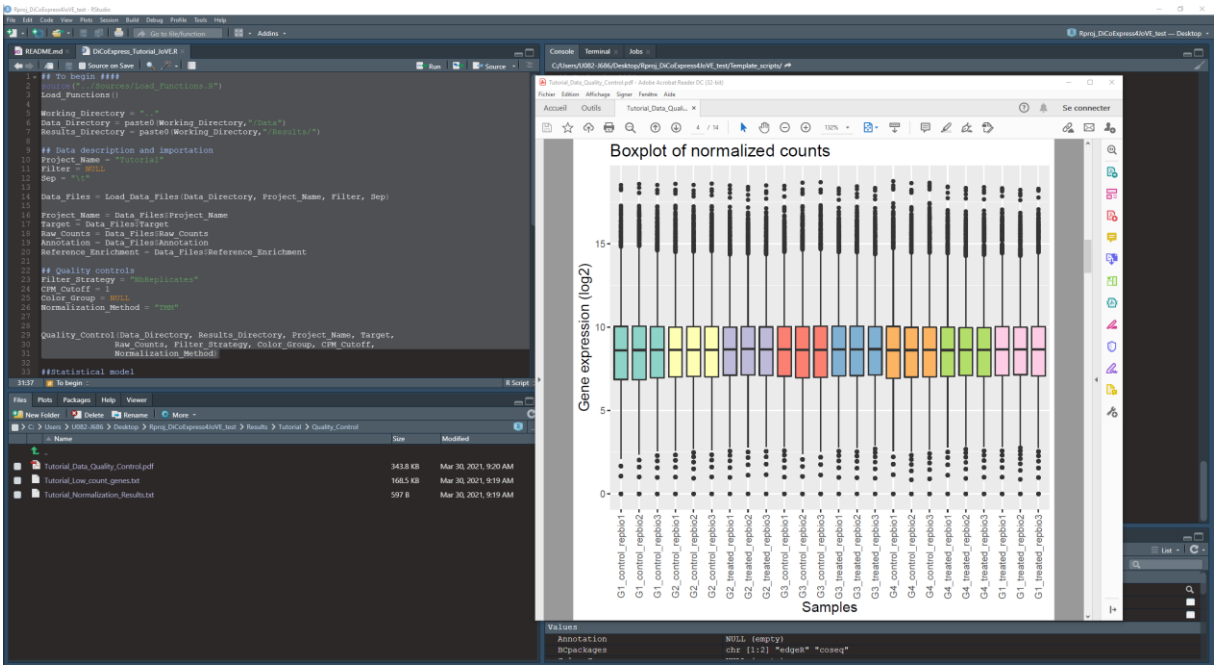
Presentation of the two mandatory files

Data analysis I : Quality control

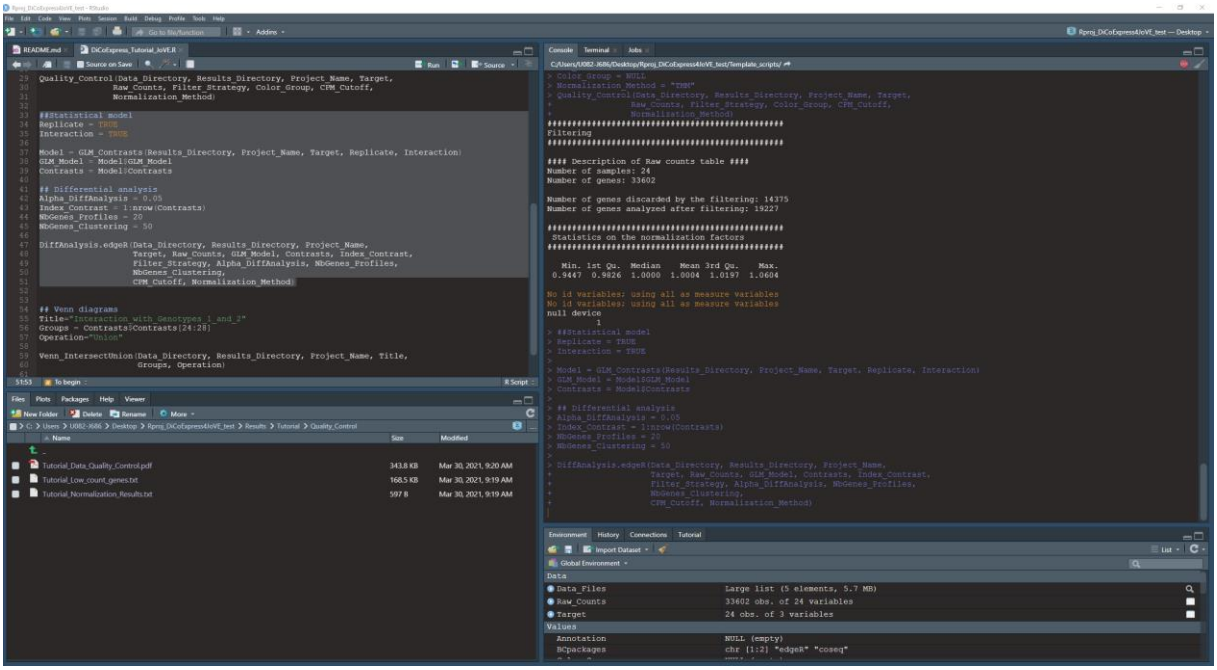


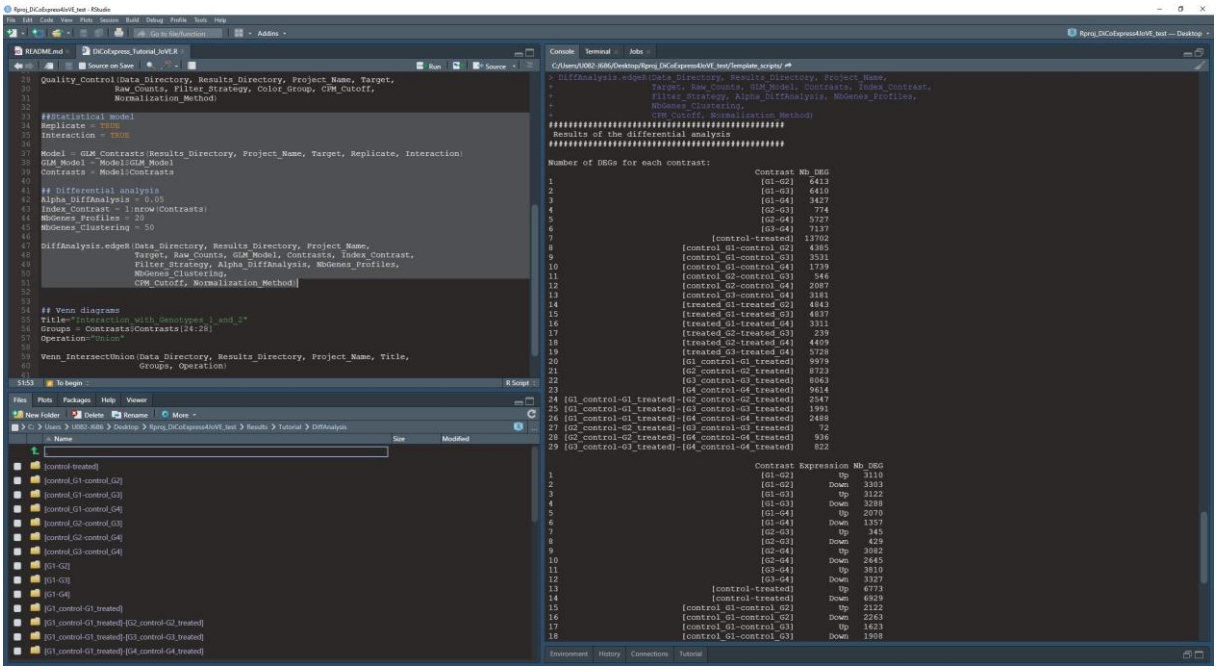
5. **Result I** : showing and commenting the QC PDF file



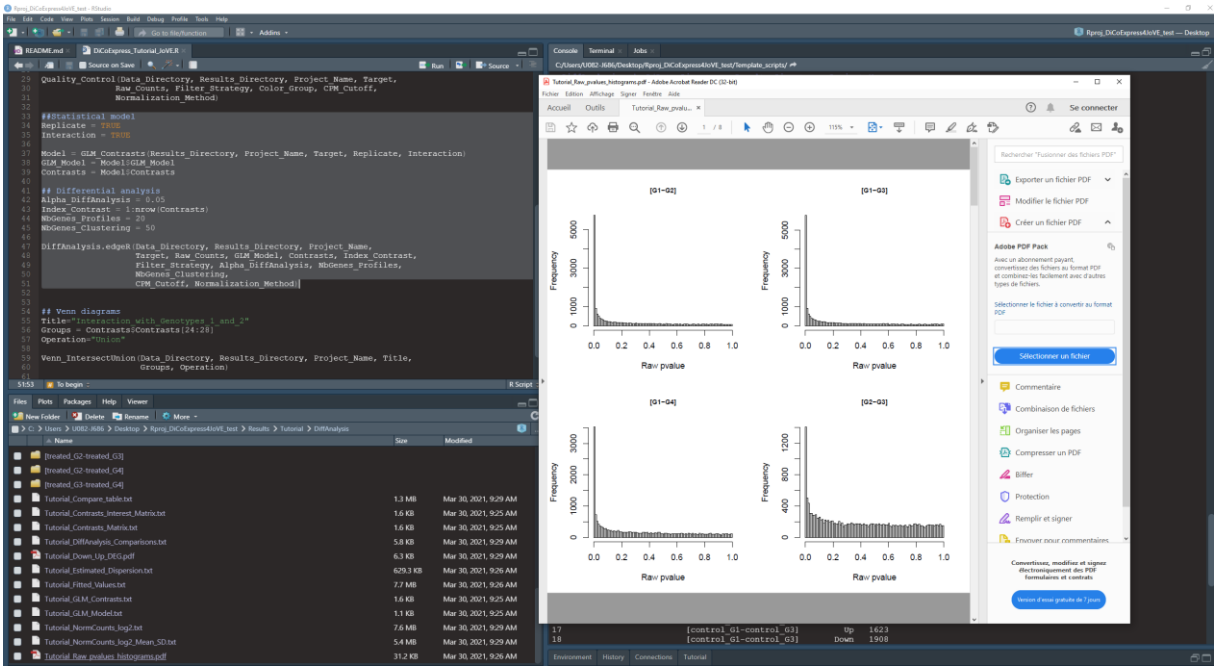


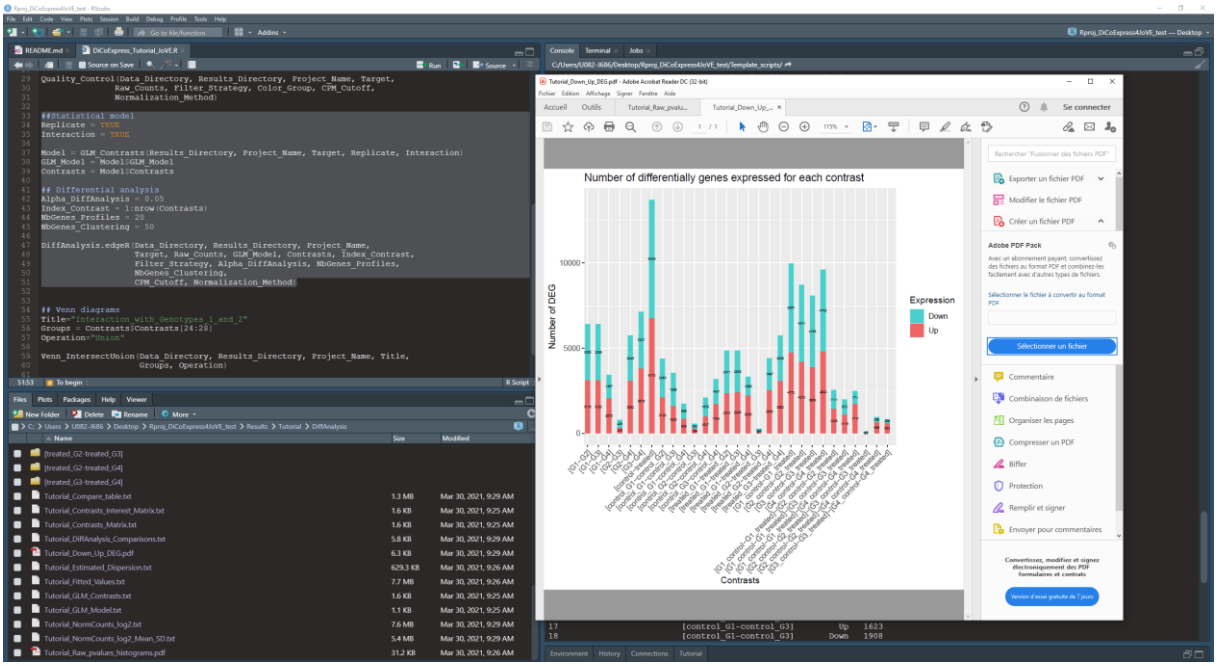
6. **Data analysis II**: Statistical modelling and differential analysis



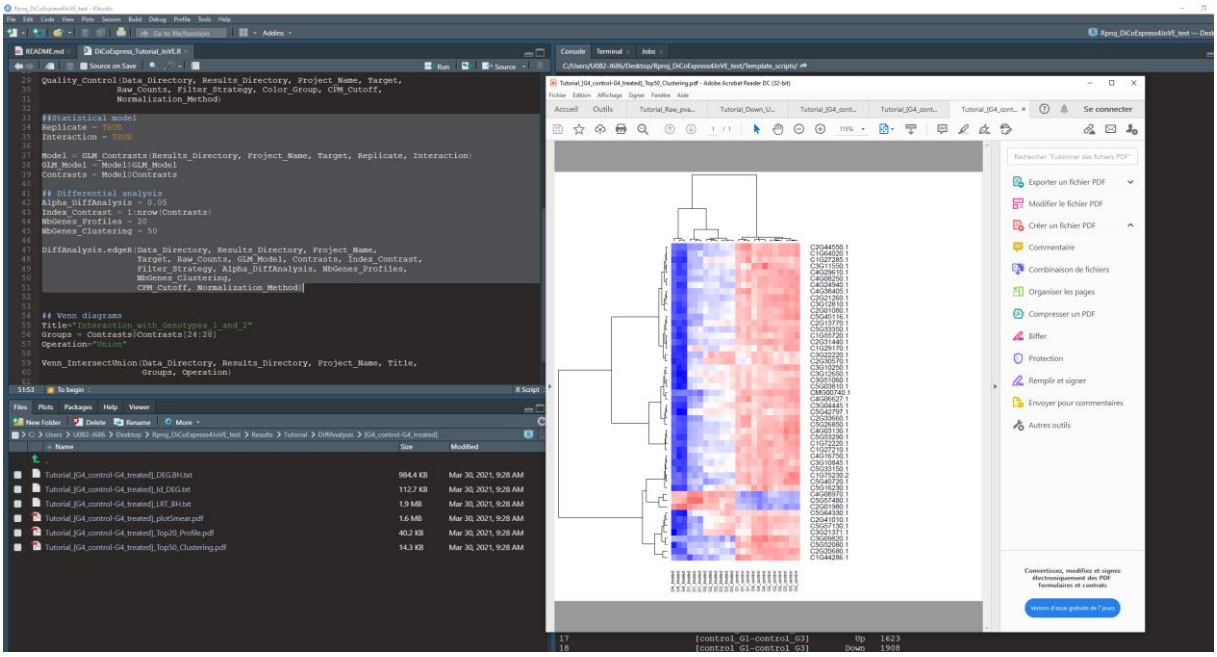


7. **Results II** : show opening it an example of differential results after showing the evaluation of p-values and commenting on it





For one contrast we'll show more of the results



8. **Data analysis III**: Enrichment of DEG lists

9. **Results III**: show and describe an enrichment table

10. **Data analysis IV**: Compare DEG lists

11. **Data analysis V**: co-expression

12. **Results V**: show co-seq output with comments on data quality check and the co-expression clusters

13. **Data analysis VI**: enrichment of the coexpressed clusters

14. **Results VI:** show an described the results

15. **Data analysis VII**: generation of log files

CONCLUSION