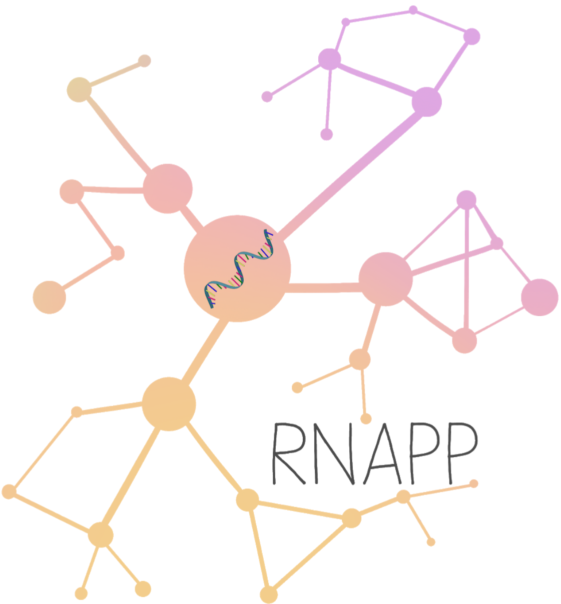
Aix-Marseille Université

M2 - Développement Logiciel et Analyses et Données

Development of an RNA-seq data analysis R-Shiny App and ATAC-seq data analysis pipeline



Margaux HAERING

Dr Bianca Habermann *supervisor*

"Computational Biology" *team*

Institut de Biologie du Développement de Marseille

School Year : 2019 - 2020

ACKNOWLEDGEMENTS

I am grateful to Mrs Bianca Habermann, Computational Biology group leader in the Institut de Biologie du Développement de Marseille (IBDM) and my internship supervisor for her welcoming in the team, all of her advices, her availability and the knowledge she shared with me.

I would like to thank Cedric Maurange, Neural Stem Cell Plasticity group leader in the IBDM and Fanny Chazal, PhD student for sharing data with me and using my app.

I thank Mr Aitor Gonzalez, associate professor and M2 DLAD supervisor for pushing me to continue my studies and look for a PhD fellowship.

I would also thank Mr Nicolas Terrapon, assistant professor and M1 DLAD supervisor for his time and his availability about PhD fellowships and making me want to go on this way.

And finally, I thank Amélie Vernale, PhD student for sharing her data with me, testing my app, and providing me feedback. Fabio Marchianò, PhD student for teaching me his tool AnnoMiner and providing me tips in R. Michaël Pierrelée, PhD student for also giving me tips in R. Paul de Boissier , PhD student, Maria Mandela Prunster, Maxime Lucas, Rikesh Jain, PhDs for their welcoming and kindness in the team.

Special thanks to my family for supporting me (and also me supporting them) and providing a comfy open workspace during this lockdown.

ABBREVIATIONS

ANR : Agence Nationale de la Recherche.

NGS : Next Generation Sequencing.

RNA-seq : RNA sequencing.

ATAC-seq : Assay for Transposase Accessible Chromatin sequencing with hightroughout.

ChIP-seq : Chromatin Immunoprecipitation sequencing.

DamID : DNA adenine methyltransferase identification.

TaDa : Targeted DamID.

GO : Gene Ontology.

TF : Transcription Factor.

TCC : Tag Count Comparison.

DE : Differential Expression.

DEG : Differentially Expressed Genes.

Log2FC : Log2 Fold-Change.

FDR : False Discovery Rate = Q-value.

NFR : Nucleosome Free Region.

PCA : Principal Components Analysis.

HMM : Hidden Markov Model.

1 - INTRODUCTION

My internship was done at the IBDM, a research laboratory primarily oriented towards developmental biology and associated pathologies. I was working at the interface of a biological team working on neuronal stem cell plasticity in *Drosophila melanogaster*, led by Cedric Maurange, and the computational biology team, led by Bianca Habermann. A joint ANR-project by the two teams aims at analyzing and integrating RNA-seq, ATAC-seq and DamID data, produced by the Maurange team during crucial steps of Drosophila neuronal differentiation.

The Maurange team is looking for elucidating the mode of action of the Chinmo/Broad transcription factor module in the developing wing in *Drosophila*. They want to understand the regulatory mechanism behind the precise transition of action of the two transcription factors Chinmo and Broad in the L3 development stage. They thus intend to decipher their mode of action on DNA, on transcription regulation, on the chromatin state and to find the targeted genes of these two factors. To address these questions, they performed RNA-seq [1] and ATAC-seq [2] at selected states of the developing fly.

The goal of my internship was the development of an analysis pipeline for RNA-seq and ATAC-seq data. Furthermore, to enable biologists to analyze their data without much assistance, I developed an R-shiny interface that allows user-friendly and easy application of the pipeline.

Biologists are typically not being trained enough in coding or data analysis to be able to use the latest technologies to analyze their data. This is where I propose this app to allow the user to analyze their data independently, to choose favorite methods easily, to generate publication-ready figures and enable them to download their data for further downstream analysis and integration or for publication.

For analyzing RNA-seq data, I developed a first server-based pipeline for quality control, read mapping to the genome and feature counting. Secondly, I built an R-based R-shiny app called RNApp for visualizing and filtering raw data thanks to multiple tools and figures. For performing a differential expression analysis using several methods of normalization and Differentially Expressed Genes (DEG) identification, I used an efficient method called Tag Count Comparison (TCC) [3] from the eponym R-package which provides a robust normalization for a robuster Differential Expression analysis. And finally, RNApp can perform GO enrichment and convert IDs between EntrezIDs, EnsemblIDs and Symbols.

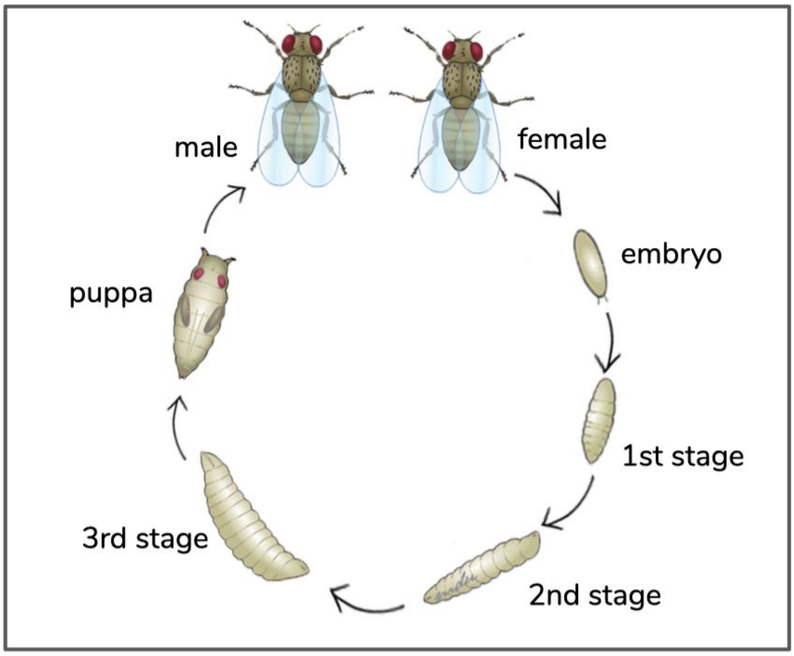


Fig 1 : D. melanogaster larval development stages.

To analyze ATAC-seq data, I developed a server-based pipeline for data treatment and analysis from the raw fastq files to the peak calling, whereby the peak calling interest is to identify regions where reads show peaks greater than the background coverage. Resulting files of this pipeline are so-called bed files that can be uploaded to the interactive genome viewer (IGV, [4]) for example.

Finally, in the Habermann team, a web-based tool for integrating peak-files with RNA-seq files was developed: AnnoMiner [5] is an annotating and enrichment tool for ChIP-seq data and as ATAC-seq data are similar to ChIP-seq data, it will be used to integrate the different datasets from the Maurange team.

2 - BIOLOGICAL CONTEXT

In development, stem cells have to decide whether to self-renew or to differentiate and in *D.melanogaster* tissues, this transition remains unclear.
Imaginal discs are structures devoted to differentiate into various parts of the insect. Ecdysone, a molting hormone regulating developmental transitions in insects is released after the larva reaches a specific, critical weight; this weight corresponds to the point in development,at which the time course to metamorphosis initiation can no longer be delayed by starvation [6].
The Maurange team is interested in the third larva stage (L3) in *Drosophila*. Larval stem cells differentiate in this stage and the development of imaginal discs can proceed to form future external structures *(Fig1)*. The Maurange team has evidence that two transcription factors, named Broad-Z1 and Chinmo, act as “pioneer transcription factors”: these are transcription factors with the ability to open the chromatin locally themselves, thus allowing other transcriptional regulators to bind and to regulate gene expression. They both act in the development of the *Drosophila* wing tissue and the team is looking for how one TF favors self-renewal (Chinmo), while the other (Broad-Z1) favors differentiation (Broad-Z1). Indeed, recent studies show that Ecdysone silences *chinmo* and activates *broad* [7].

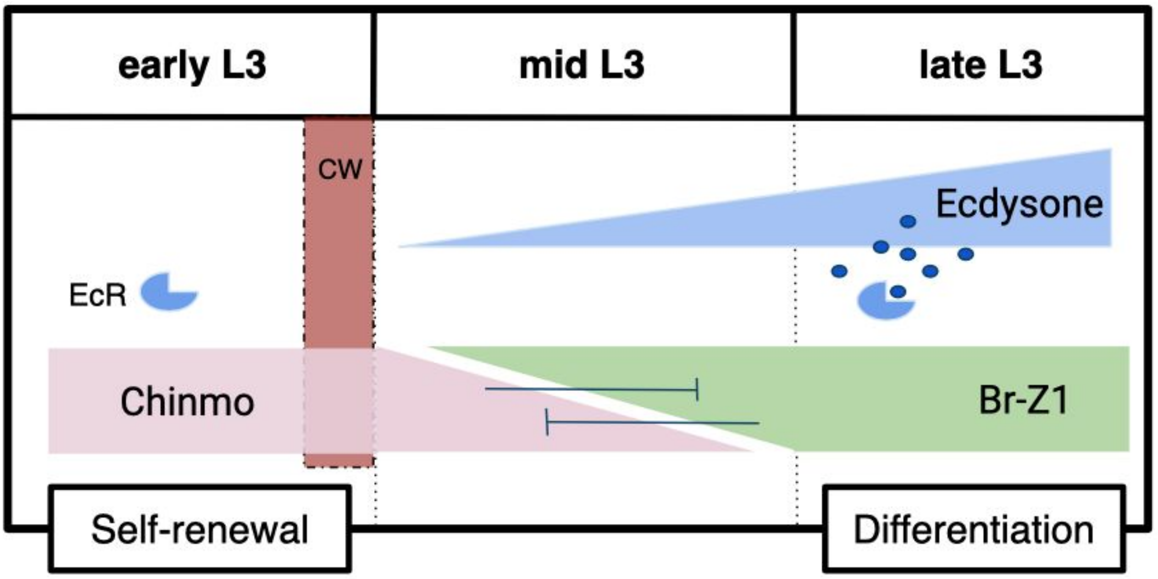


Fig 2 : Within the third larval stage (L3), Ecdysone coordinates self-renewal and differentiation across the Chinmo/Broad-Z1 bistable loop. Early,mid,late L3 are phases in the L3 larval development stage. CW :Critical Weight and EcR : Ecdysone Receptor.

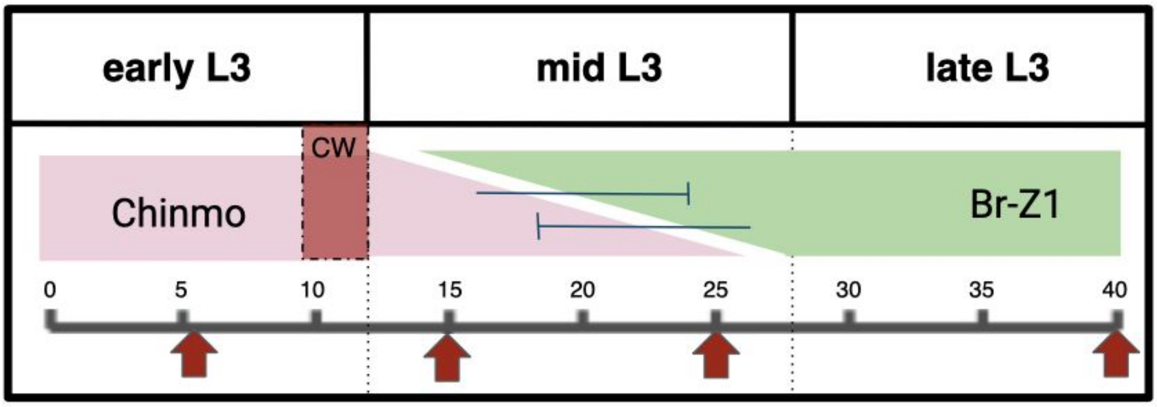


Fig 3 : Dissection time points for RNA isolation. eL3+6H, eL3+15H, eL3+25H and WL3 for Wand. CW :Critical Weight and EcR : Ecdysone Receptor.

Chinmo is a member of the Zinc Finger and BTB protein family (ZBTB) family. It has been shown to have a role in the regeneration of the wing discs; more precisely, it is up-regulated in the blastema [8] in the early L3 stage in Drosophila. The blastema (or Regeneration Bud) contains undifferentiated proliferating cells that have the capability of growth and regeneration and, during development, of developing into organs or other body parts.
The *broad* gene is coding for the Broad-Z1 transcription factor from the ZBTB framily and is an early target of ecdysone. *Broad* promotes differentiation of epithelial cells and is activated by the release of ecdysone.

Chinmo and Broad-Z1 form a bistable loop through cross-repression coordinated by ecdysone *(Fig 2)*. After the release of ecdysone, the regeneration process stabilized by Chinmo is no longer active. Broad-Z1 takes over and by repressing Chinmo, going back to a self-renewal state is no longer possible. If tissue damage occurs in the Chinmo phase (early L3), it can regenerate with the help of the *wingless* and *Wnt6* gene. This process is however limited by the transition from the early L3 to the late L3 stage [9].

The Maurange team wants to understand the molecular details of the Chinmo/Broad transition. They want to know which target genes are activated by the two transcription factors. To this end, they did RNA sequencing using Illumina sequencing at four selected time points during the L3 stage of larval development.

The very first and last time points collected, eL3+6H and WL3 can be viewed as controls, as either only Chinmo or only Broad-Z1 is expressed, representing the self-renewal and differentiation stage, respectively. These stages are nonetheless useful to identify genes corresponding to these two states. The two timepoints during the transition represent the interesting states in the way that it will show the transition in gene expression between the first and last cluster, identifying precisely which genes are active during the transition period. These time-points are selected at the beginning and the end of the transition *(Fig 3)*.

A complementary approach to study this process is to analyze the changes in chromatin state during the transition. The nucleosome is the elementary unit of chromatin composed of DNA wrapped around an eight subunit histone core. The genomes of eukaryotes are packaged in such a way into chromatin. In the Nucleosome Free Regions (NFRs), DNA is more accessible due to the absence of nucleosomes. Consequently transcription factors can bind DNA more easily, controlling gene-expression which is for instance critical in the establishment of cellular identity during development. Linked to the regulation of the transcription studied previously with RNA-seq, ATAC-seq will shed light on the changing of the chromatin state and we will identify the genes responsible for opening and closing the chromatin.

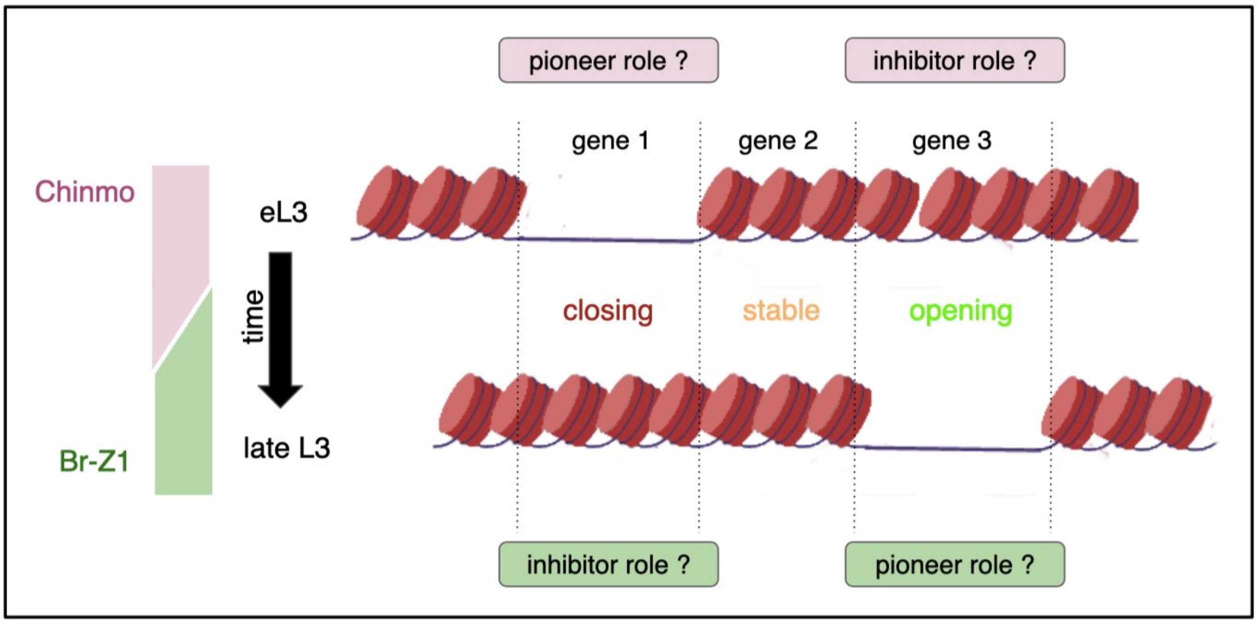


Fig 4 : Chromatin dynamic supposition over time. Chinmo and Broad-Z1 (Br-Z1) having a pioneer role in opening chromatin in early L2 stage (eL3) and late L3 stage, respectively; and an inhibitor role closing chromatin reciprocally.

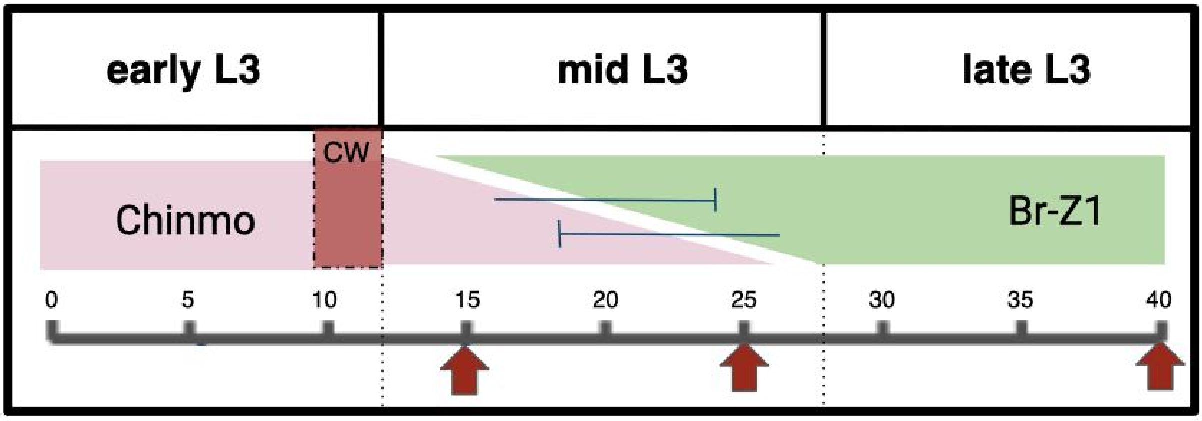


Fig 5 : ATAC-seq dissection time points: eL3+15H, eL3+25H and WL3 for Wand. CW : Critical Weight.

ATAC-seq allows to find these NFRs by sequencing specifically DNA from previously isolated NRFs. The Maurange team hypothesizes that Chinmo binding sites are majorly positioned on closed chromatin sites and wants to verify this hypothesis. More generally, the aim is to check if any changes of chromatin state occur over time from early L3 to late L3 stages and more interestingly, which chromatin state corresponds to self-renewal and which one to differentiation. By performing ATAC-seq, they can address the question, whether Chinmo and Broad-Z1 are “pioneer transcription factors” (*Fig 4*).

ATAC-seq is currently in progress, whereby 3 time points were selected : eL3+15H and eL3+25H and the last one in the late L3, WL3 (*Fig 5*). The first time-point corresponds to the chimno-state of self-renewal. As wing imaginal discs are too small in the early L3 stage and not enough biological material can be collected to perform ATAC-seq, a later time-point (eL3+15H) was chosen. The second one will show the transition between the *chinmo*+ and *broad*+, in other words, between self-renewal and differentiation; the last time-point will show the state of the chromatin in the differentiation stage.

3 - MATERIAL & METHODS

RNA sequencing

RNA-seq is a technology based on the Next Generation Sequencing (NGS) allowing robust quantification of transcripts in a cell population or a tissue [1]. It allows the study of the regulation of transcription in a certain condition, at a certain time point or even across time.

RNA is isolated depending on the type of RNA to profile (mRNA, total RNA, miRNA, ncRNA) and a library is constructed. It can be single-end (stranded or not), where only one end of the fragments are sequenced or paired-end where both ends are sequenced which facilitates alignment allowing a better reconstruction of the transcripts. Each library is then sequenced and raw data are provided in the fastq format. Each base is given a Sanger quality score. The Sanger score is a measure of the quality per base call based on the probability p of error: $QSanger = -10 * log_{10}(p)$.

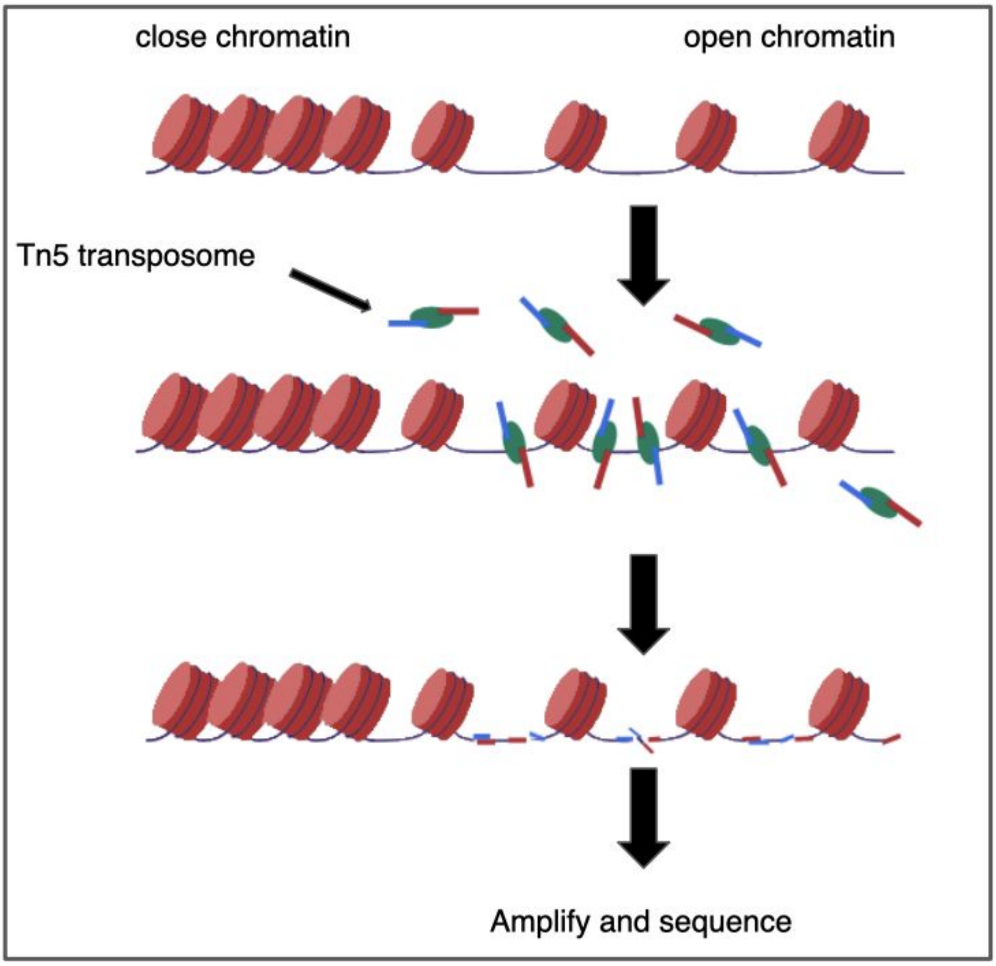


Fig 6 : Schematic ATAC-seq library preparation protocol. Tn5 transposases allow the placement of adapters to tagged chromatin fragments in open chromatin regions.

ATAC sequencing

ATAC-seq is a NGS technology based on preferential integration of a transposon in open chromatin to capture open chromatin sites. It has the ability to determine nucleosome positioning and to identify transcription factor footprints, as binding sites are protected from transposon insertion [2]. The library is constructed using the Tn5 transposase and adapters are loaded onto fragments. Next, chromatin is fragmented, allowing integration of adapters into open chromatin sites (*Fig 6*).

FastQC

A quality control tool aiming to check quality of raw sequence data coming from NGS sequencing [10]

cutadapt

A trimming command line tool to remove adapters from NGS reads, supporting 454, Illumina and SOLiD sequencing data [11]. cutadapt does not use alignment scores but unit costs where mismatches, insertions or deletions are one error. This method allows to submit a maximum error rate as a single parameter to the algorithm. Adapter sequences are provided by the user and cutadapt searches and removes best matches to the sequence. A minimum overlap can also be provided to minimize random hits; the tool has many options to choose from.

featureCounts

A software to count reads located on genomic features using chromosome hashing. featureCounts works with single or paired-end reads. It also provides a lot of options. [12]

STAR

An alignment and mapping software to map sequencing reads to their reference genome. STAR is especially designed for RNA-seq and is very fast and accurate. It is not an extension of a short-read DNA mapper like many RNA-seq mapping software are and provides an option to inform about possible splice junctions [13].

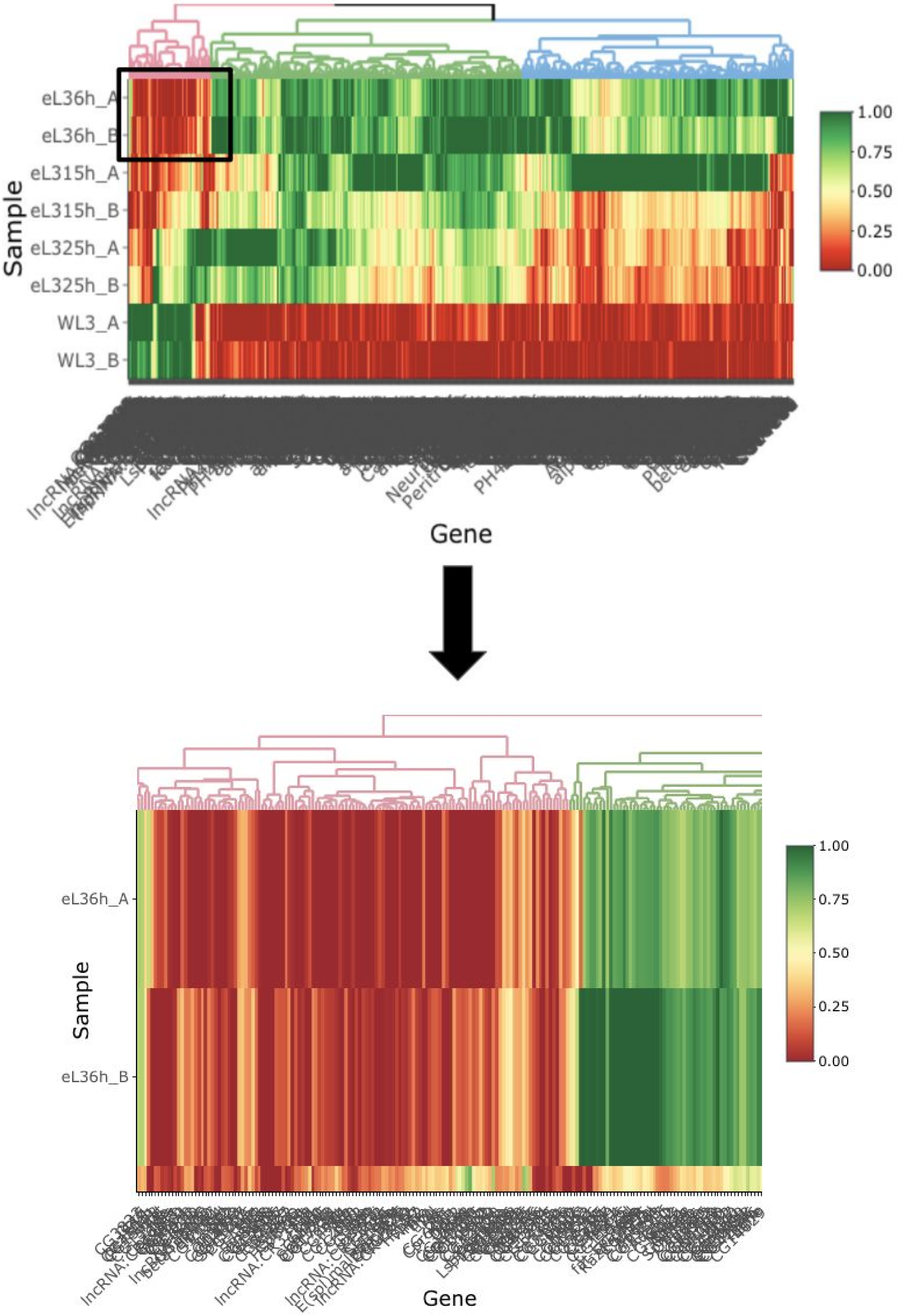


Fig 7 : Illustration of the Plotly R-package, on zooming in. This example is showing real data and 9081 genes in the global heatmap view.

Bowtie2

Bowtie2 is also an alignment and mapping software that is a full-text minute index-based aligner, meaning it does not use a genome index to permit gapped alignment. It is based on the bidirectional Burrows-Wheeler transform [14] and is typically used for mapping ChIP-seq or other short read data where no splicing is expected.

MACS2

MACS2 is a peak calling tool identifying genomic regions significantly enriched designed for ChIP-seq data. It analyzes short reads mapped to the genome to localize genome-wide transcription factor binding sites and can also be used to identify NFRs coming from ATAC-seq data. MACS2 removes repeated reads, adjusts read positions, calculates peak enrichments and finally estimates the FDR [15] [16].
Output files of MACS2 correspond to a standard BED file and narrowPeaks specific fields. We can find a measurement of overall enrichment of the region, a statistical significance (in form of a p-value), a false-discovery rate (FDR), and the positions of the peaks, which is the point source called the peak.

R

A statistical software suite widely used for statistical analysis; It is known for its efficiency, its ease-of-usage, and many Apps downloadable as Bioconductor packages which provide many biological analysis tools.

Plotly R-package

Plotly is an R-package to create, among others, interactive figures, which are downloadable, and which are easy to zoom in and out, rescale, and which can be opened in a separate window (*Fig 7*).

Shiny R-package

Shiny-R is a package to build interactive R-apps, in which it is possible to include figures, tables, Rmarkdowns, etc. It offers a wide range of customizations.

GO enrichment

Enrichr [17] is a web-based tool for gene set enrichment analysis and returns any enrichment of common annotated biological features. Enrichr R-package provides an interface to all Enrichr databases.

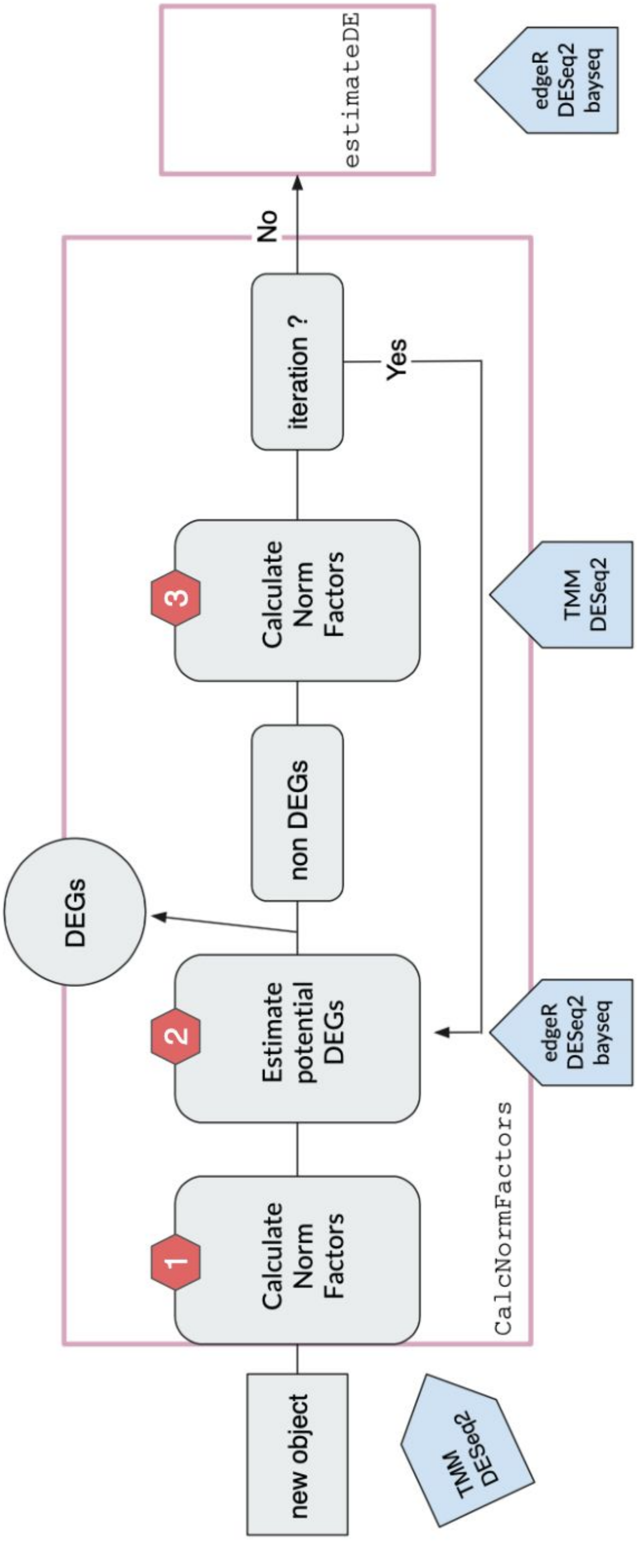


Fig 8 : TCC calculation process. The CalcNormFactors function allows to remove potential DEGs in step 2 before the normalization in step 3 to have robust normalization factors. Also these two steps can be repeated in a loop to obtain a more accurate list of DEGs and normalization factors. estimatedDE function performs a DE analysis with the chosen method on the flagged DEGs with a given FDR threshold.

TCC R-package and calculations

The TCC R-package [3] optimises the normalization and DEG identification step of RNA-seq analysis. It offers the opportunity in 2 R-functions to realize these steps and includes the possibility to suppress potential differentially expressed genes before normalization. The method is called Differentially Expressed Genes Elimination Strategy (DEGES) and is performed in the `CalcNormFactors` R-function. It calculates in a first step every normalization factor, entering a loop of normalizations and DEG identifications; it then flags potential DEGs in a second step and removes them for the next normalization factor calculation steps. The third step recalculates normalization factors to make them more accurate. If iteration is chosen, it redoes steps 2 and 3 in a loop as long as the iteration factor given. This iterative normalization process allows a robust normalization and accurate normalization factors.
The second R-function `estimateDE` performs the DE analysis over all flagged DEGs with the chosen method DESeq2, edgeR or bayseq (*Fig 8*).

The result of the calculation is a variable gathering Log2FC, BaseMean, P-value, FDR and also a ranked list where the most deregulated genes are top ranked and non-DEGs are bottom ranked with respect to their p-value.

Normalization methods provided are TMM [18] Trimmed Means of M values, a global scaling method and DESeq2 [19] which uses the median of ratios method and estimates of dispersion and logarithmic fold changes.
DE analysis methods that can be chosen from are DESeq2 [19] that uses shrinkage estimation for dispersions and fold changes; EdgeR [20], an overdispersed Poisson model used to account for both biological and technical variability. And finally, bayseq [21] uses an empirical Bayesian approach to define patterns of differential expression.
Finally, the authors of TCC developed the TbT method (TMM-bayseq-TMM) that they recommend for two groups with or without replicates and has proven its efficiency on calculation time. The goal of developing such a package with a robust normalization method was motivated by the fact that normalization has a greater impact on the ranking of genes than the DEG identification method [22].

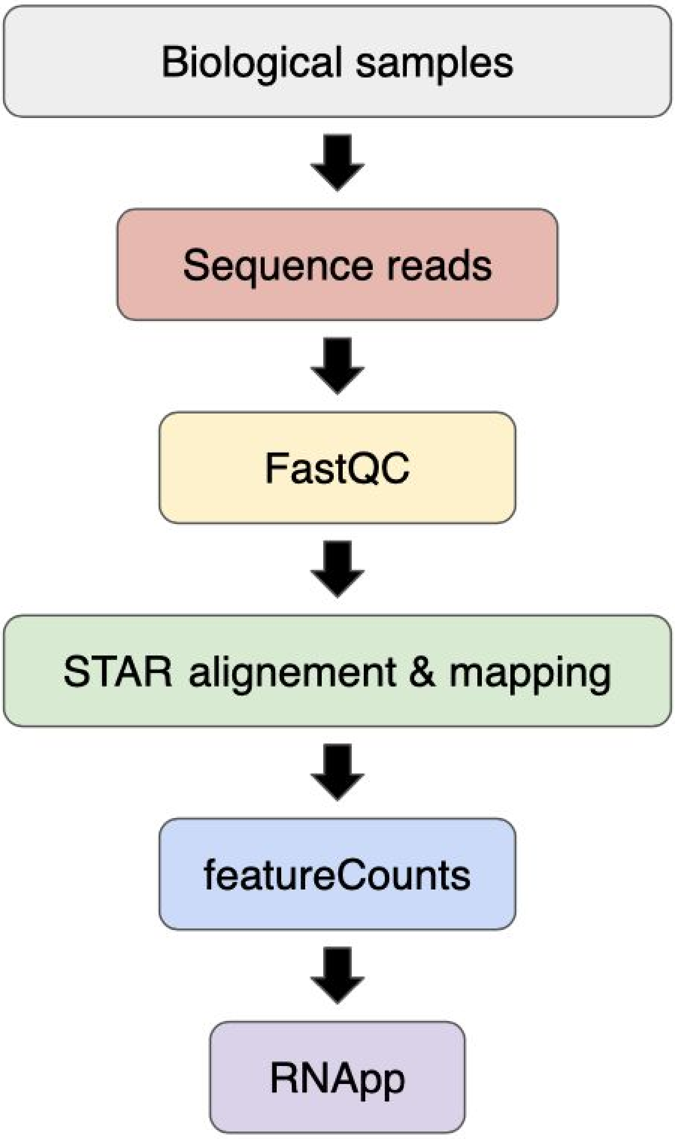


Fig 8 : Preprocessing analysis pipeline for RNA-seq data.

4 - RESULTS

4.1 - RNA-seq analysis

Biologists need to be free in their choices of analysis tools to be efficient, and to get the best results from an RNA-seq experiment. The development of a user-friendly, interactive and comprehensive analysis-pipeline was in our minds the best solution.
RNApp is an interactive R-Shiny app that provides different methods and visualization choices. Starting from a count matrix, it contains all analysis steps starting from normalization and ending with functional enrichment.

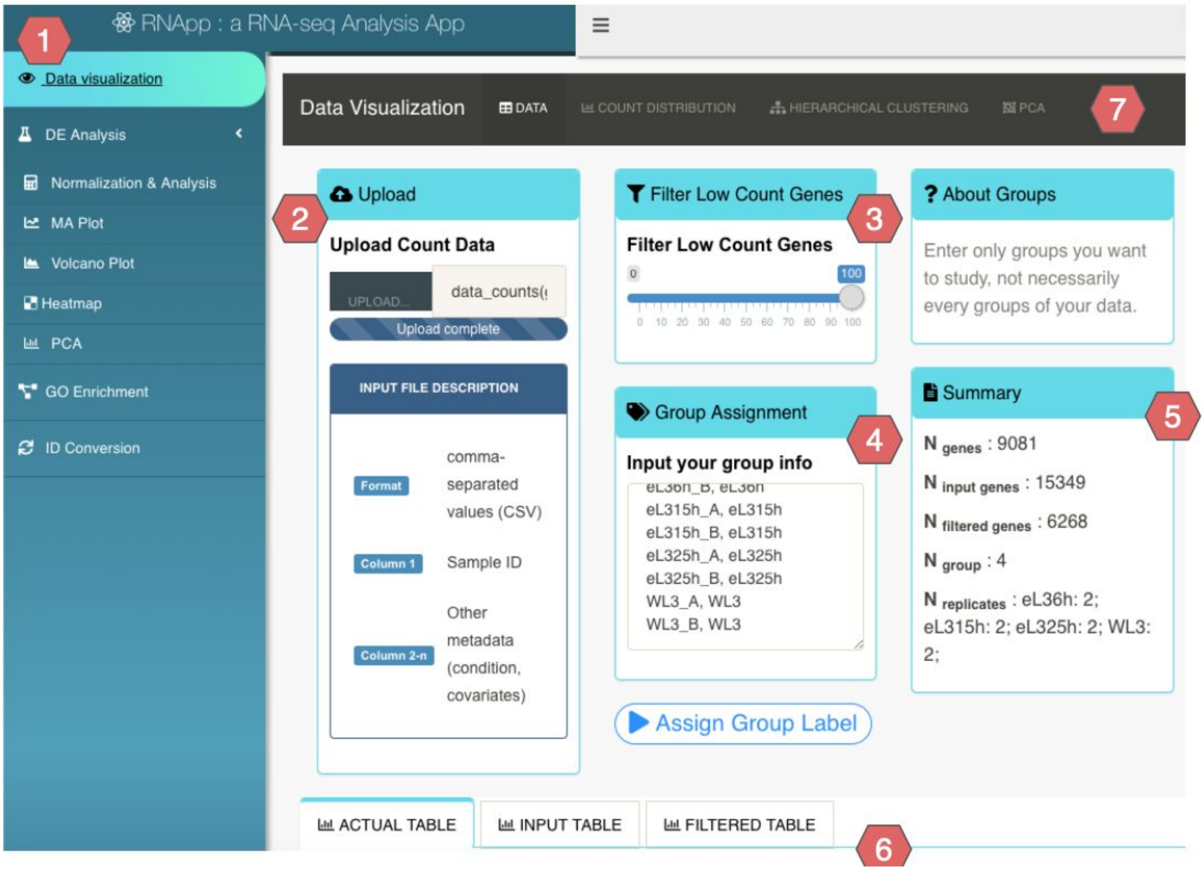
4.1.1 - Preprocessing, read mapping and feature counting

Quality control, read mapping to the genome and feature counting needs to remain command-line based as it needs a decent server with sufficient disk space, memory and a sufficiently fast processor to run, where it still takes a few hours, depending on the server configuration and the number of samples to analyse.
The pipeline I developed for these steps is very simple, providing a classic quality control with FastQC for every single sample and read mapping to the genome with STAR [13] (*Fig 8)*.

FastQC analyses the quality of the reads. Low-quality reads or bases can then be trimmed, if wanted. For trimming, Trimmomatic [23] can be used for example, which is a trimming tool for Illumina reads. However, it has been hypothesized that trimming alters gene expression estimates and can thus lead to biased results [24]. Therefore, we decided not to trim reads but to proceed with raw reads.

STAR (Spliced Transcripts Alignements to a Reference) is a fast alignment and mapping tool especially designed for RNA-seq data. It is capable of running on several threads on a multicore system. Read mapping on our system for the 8 samples I had (4 time-points with two replicates each) took approximately 4 hours (*Fig 3*).

FeatureCounts was used to assign reads to genomic features and obtain a count matrix of expression that will be the input file for RNApp.



Upload Count Data

UPLOAD... data_counts(

Upload complete

INPUT FILE DESCRIPTION

Format	comma-separated values (CSV)
Column 1	Sample ID
Column 2-n	Other metadata (condition, covariates)

Filter Low Count Genes

0 100

0 10 20 30 40 50 60 70 80 90 100

Enter only groups you want to study, not necessarily every groups of your data.

Input your group info

eL36h_B, eL36h
eL315h_A, eL315h
eL315h_B, eL315h
eL325h_A, eL325h
eL325h_B, eL325h
WL3_A, WL3
WL3_B, WL3

N genes : 9081

N input genes : 15349

N filtered genes : 6268

N group : 4

N replicates : eL36h: 2; eL315h: 2; eL325h: 2; WL3: 2;

Fig 9 : Opening page of RNApp. Sidebar menu of all tools the app contains (1). The upload box (2). The filter of low count genes; the summary is updated in real time after filtering. (3). Group assignment box (4). Summary of the input data (5). Tables of the input data, filtered and full input, as well as removed genes (6). Navigation bar of visualization of raw data (7).

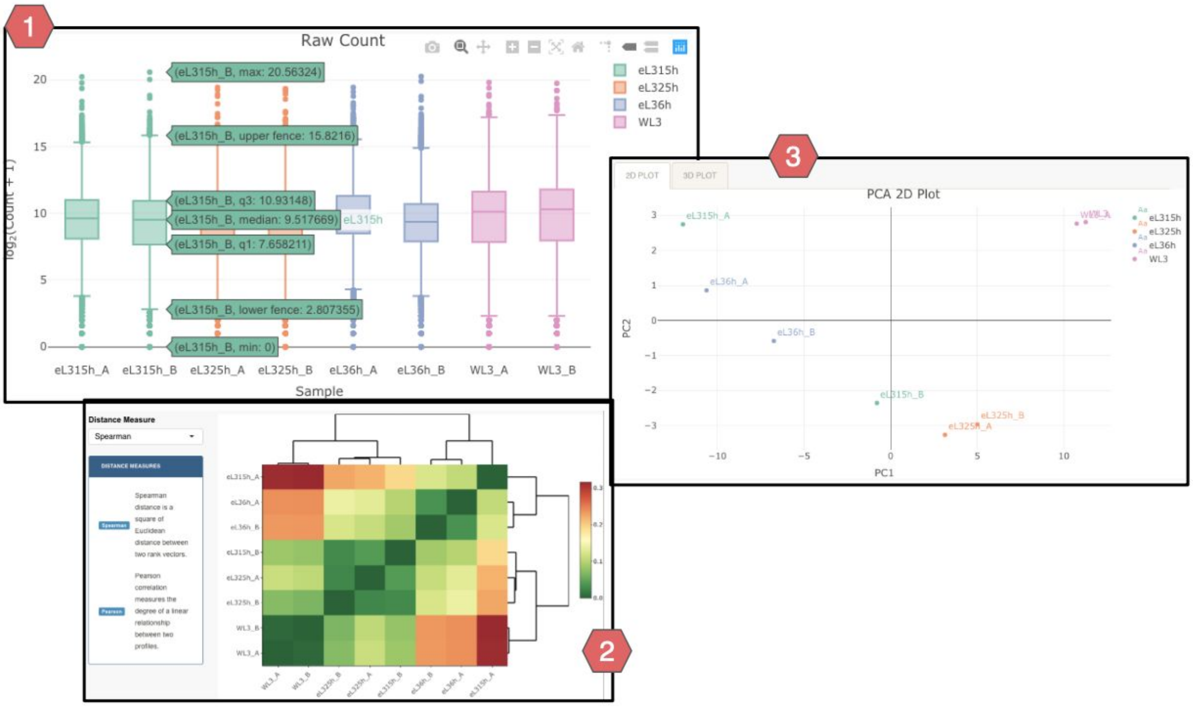


Fig 10 : Visualization of raw data including a box plot with statistics (1). A hierarchical clustering where the method for calculating the distance can be chosen (2). A 2D and 3D PCA clustering of samples (3).

4.1.2 - RNApp

R provides a lot of tools and is easy to use. Using Shiny, an R package to build and design interactive applications which are very easy to use, was an obviousness choice. It can either be created as a single app.R file; or as separate files, where one deals with the user interface (ui-files) and the other with the backend (server-files). I decided to employ the second strategy, as user interface and server-files allow to separate what is visible to the user and what is in the back-end of the app. They also make it easy to navigate between the user interface and the code, as every step has its own ui and server file. For example to build an MA plot, ui-maplot.R is referencing the shiny part, so the part that is visible in the app, and server-maplot.R the process of generating the MA plot.

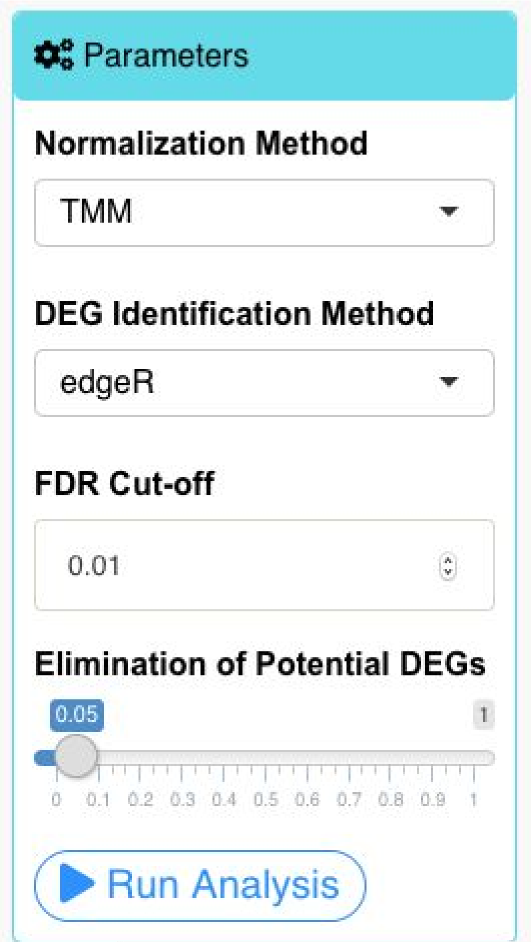
RNApp is built with 22 files including 3 Rmarkdowns of information, explaining the analysis steps, the enrichment part and the ID conversion part. As I decided to make the dashboard customizable, I used a theme made by nik01010 on GitHub that was easy to install thanks to the `devtools` R-package. Finally, I used the `plotly` R-package to build all figures generated in the app, to make it really user friendly, as well as allow downloadable images. Code is available on GitLab : https://gitlab.com/habermann_lab/rna-seq-analysis-app

4.1.2.1 - Raw count data

Opening RNApp, we can find a sidebar menu on the left to have easy access to all tools at the starting page (*Fig 9*): uploading a count matrix, filtering to remove low count genes if wanted, and a group assignment to assign replicates and conditions. This step is crucial to tell the app what to compare. Initially, the input count matrix can contain many conditions (groups), but the user might want to compare only two particular conditions, for example. The user just has to assign the desired groups with their replicates in the right order (from the original count matrix) and the analysis will be done on the chosen conditions only.

This view also provides a summary of the input genes and groups with a filtered table containing the chosen genes to do further analysis, a table containing removed genes due to the filtering of low count genes, and the full input table.

Finally, we can find visualization tools for raw data (*Fig 10*), including a box plot of count distribution with a hover functionality so that we can obtain all statistical data for each sample (`plot_ly` function with argument `type = "box"`). It also provides a heatmap based on groups to perform a hierarchical clustering between groups. Here, the user can choose the distance measure between Spearman distance and Pearson correlation. The heatmap is realized with the `heatmaply` function from the eponym R-package that performs heatmap building with the plotly package.



Parameters

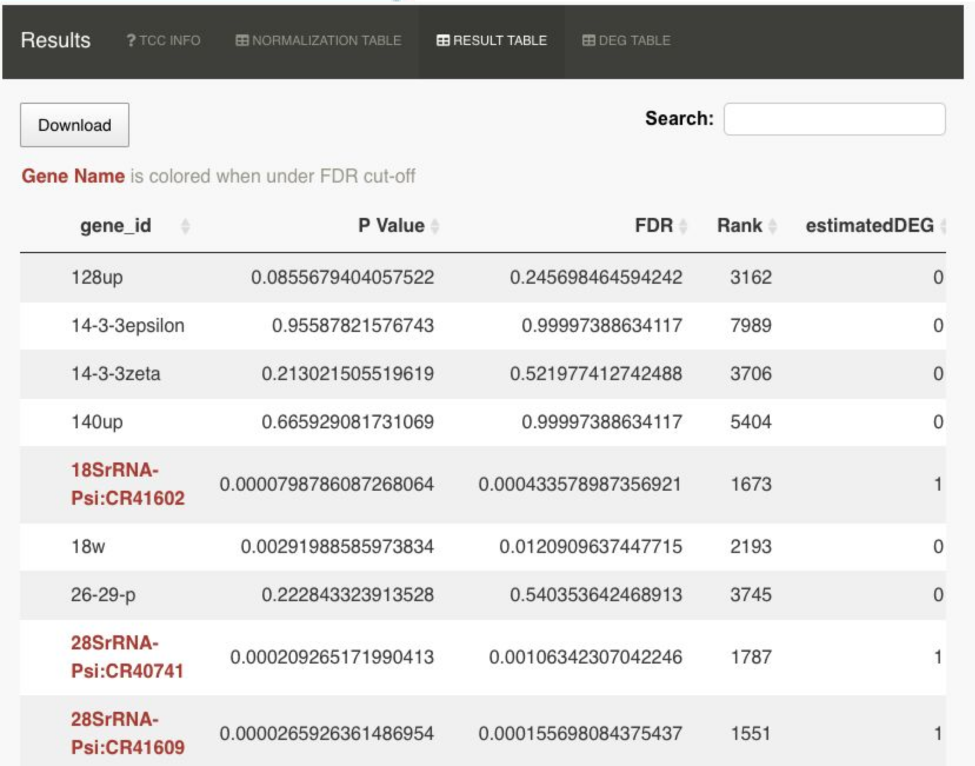
Normalization Method: TMM

DEG Identification Method: edgeR

FDR Cut-off: 0.01

Elimination of Potential DEGs: 0.05

Fig 11 : Normalization - DEG identification step parameters. Normalization methods are TMM and DESeq2. DEG identification methods are edgeR, DESeq2 and bayseq. The FDR cutoff is freely selectable as are the proportion of potential DEGs,; the default value is 0.05.



Gene Name is colored when under FDR cut-off

gene_id	P Value	FDR	Rank	estimatedDEG
128up	0.0855679404057522	0.245698464594242	3162	0
14-3-3epsilon	0.95587821576743	0.99997388634117	7989	0
14-3-3zeta	0.213021505519619	0.521977412742488	3706	0
140up	0.665929081731069	0.99997388634117	5404	0
18SrRNA-Psi:CR41602	0.0000798786087268064	0.000433578987356921	1673	1
18w	0.00291988585973834	0.0120909637447715	2193	0
26-29-p	0.222843323913528	0.540353642468913	3745	0
28SrRNA-Psi:CR40741	0.000209265171990413	0.00106342307042246	1787	1
28SrRNA-Psi:CR41609	0.0000265926361486954	0.000155698084375437	1551	1

Fig 12 : DE Analysis result table. It provides the normalization values on a right side slide.

Finally, a 2D and 3D Principal Components Analysis (PCA) of the groups done with the `prcomp` function from the `stats` R-package is shown.

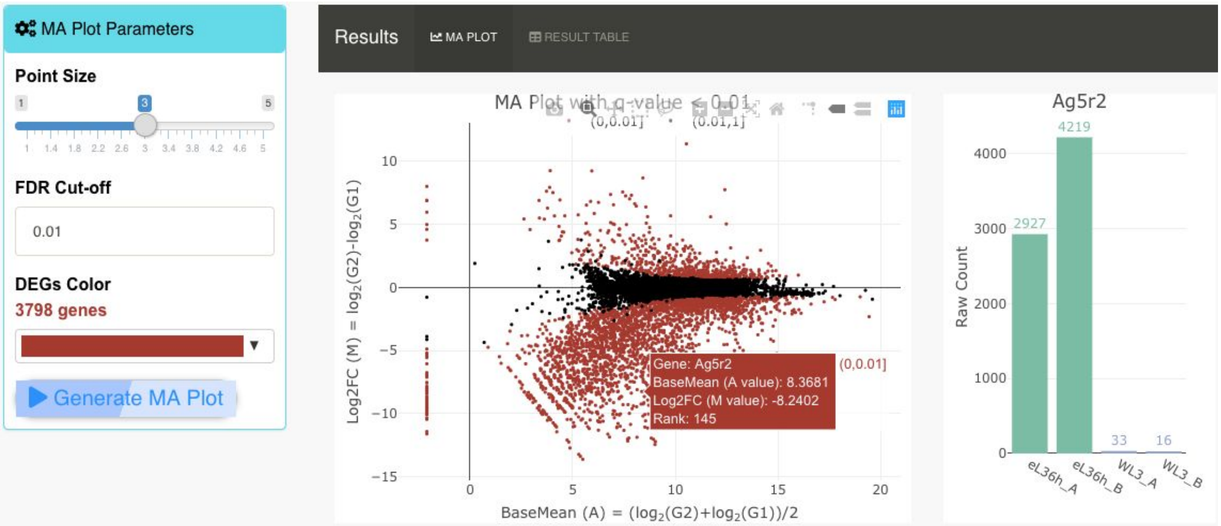
4.1.2.2 - Normalization and differential expression analysis

The core steps of an RNA-seq analysis are the normalization and the DEG identification. DESeq2 is a very good and useful analysis package and is often used in RNA-seq analysis. However, I wanted to offer the user the opportunity to choose from several possible methods. I wanted to include DESeq2 and edgeR as analysis packages in RNApp and allow the user to choose between them. Thus, I decided to implement the TCC R-package. This package offers TMM and DESeq2 as normalization methods; and edgeR, DESeq2 and bayseq as DEG identification methods. It therefore was the ideal choice for implementing several methods at once.
The package is user-friendly and uses two functions to perform normalization and DEG identification, using any of the included methods. For example if both methods chosen are DESeq2, TCC is performing the analysis using the two individual steps from the `DESeq2` package independently, DESeq2 normalization and DESeq2 differential expression analysis.

In RNApp, I added a Rmarkdown to explain how the package works, as the user has to choose both methods, normalization and DEG identification; furthermore, an FDR cut-off and more importantly a value of potential DEG elimination (*Fig 11*) has to be chosen. For users not familiar with TCC, this terminology can be misleading, because an “elimination” would suggest removing this proportion of potential DEGs. But, as introduced previously, the normalization procedure used by TCC removes a proportion of potential DEGs during the loop of normalization as part of a robust method of normalization. These removed DEGs are stored in the results awaiting the rest of the iterative normalization procedure to be completed. The default value of potential DEGs elimination is 0.05 with respect to the p-value.

I decided to set the iterations to three by default to provide user-friendliness and to avoid misunderstanding of this step in RNApp. This setting should be discussed and maybe open to a choice.

After the TCC calculation is done, a result table is made available offering all resulting statistics including Log2FC, BaseMean (these two only in the case of comparing two conditions), P-value, FDR, and the rank (*Fig 12*). The ranking shows the most deregulated genes top ranked and non DEGs bottom ranked. It also provides the normalization expression values. Next to the complete result table, the user can also download only normalized read counts or only the DEGs.



MA Plot Parameters

Point Size: 3

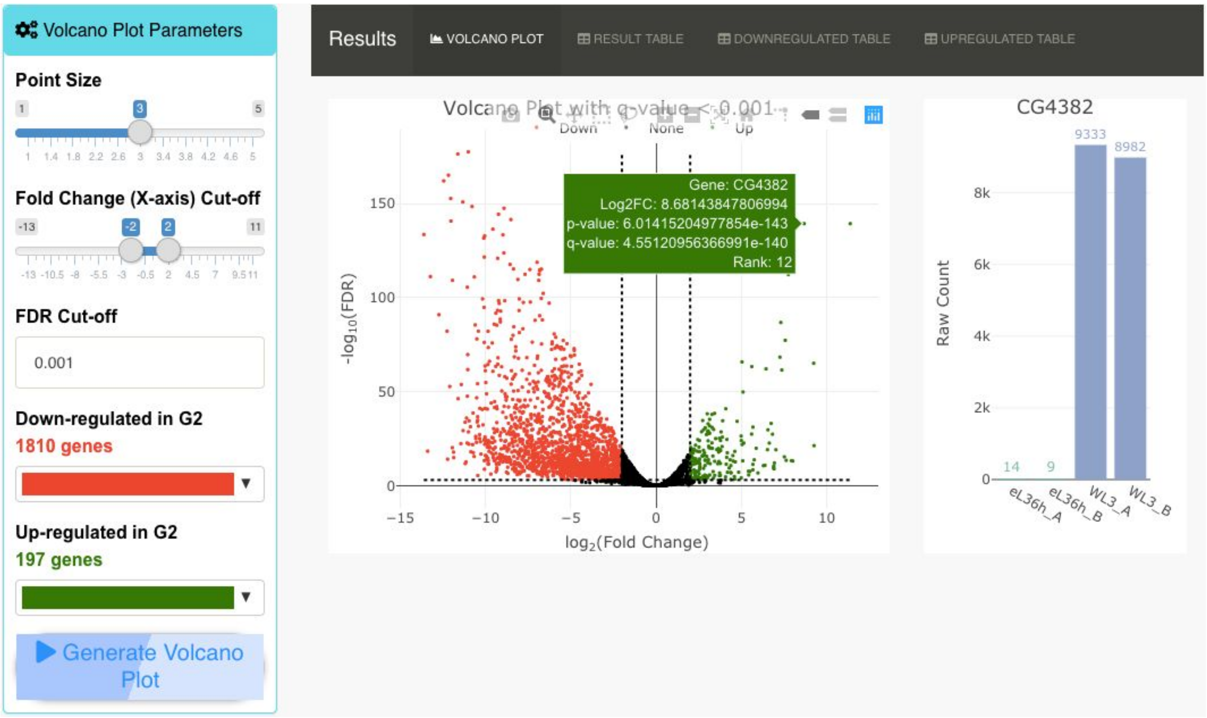
FDR Cut-off: 0.01

DEGs Color: 3798 genes

Results | MA PLOT | RESULT TABLE

Gene: Ag5r2
BaseMean (A value): 8.3681
Log2FC (M value): -8.2402
Rank: 145

Fig 13 : MA plot with its associated bar chart of raw expression, when hovering over a gene. The color of the DEGs is selectable by the user just like the size of the points, as well as a selection by an FDR cut-off.



Volcano Plot Parameters

Point Size: 3

Fold Change (X-axis) Cut-off: -2 to 2

FDR Cut-off: 0.001

Down-regulated in G2: 1810 genes

Up-regulated in G2: 197 genes

Results | VOLCANO PLOT | RESULT TABLE | DOWNREGULATED TABLE | UPREGULATED TABLE

Gene: CG4382
Log2FC: 8.68143847806994
p-value: 6.01415204977854e-143
q-value: 4.55120956366991e-140
Rank: 12

Fig 14 : Volcano Plot with its associated bar chart of raw expression when hovering over a gene. Point size, Fold Change cut-off, FDR cut-off and colors of DEGs are customizable.

4.1.2.3 - Visualization of results from the DE analysis

To visualize results from DE analysis, I added several plots.
For comparative analysis of two conditions, a MA plot and a Volcano plot are available. The MA plot relates the M value (Log2FC) to the Log2(A-value) (BaseMean), which is the mean expression. These two values are calculated in the course of the DE analysis. The `TCC` R-package offers a `plot.TCC` function, producing a MA plot. However, I wanted my figures to be done with the `plotly` R-package. So I produced my own MA plot with `plotly`. Thanks to the argument "`key`" of the `plot_ly` function, I was able to create an associated bar chart. Thanks to `plot_ly`, the user can see the raw read counts of a gene in a specific sample, when the gene in question is hovered over in the MA plot. It also shows M and A values and the rank of the gene on the MA plot (*Fig 13*).
A result table is also available for download, if the FDR cut-off is reduced to show the MA plot. If it is not changed, it is the same table produced and made available already during the DE analysis.

For comparative analysis between two conditions, RNApp also presents a Volcano plot in the same style as the MA plot, where a bar chart is associated with raw count data. The Volcano Plot represents the -log10(FDR) with respect to the log2FC and I also used `plot_ly` to produce it. When a gene is hovered over, its log2FC, P-value, FDR and rank is given. Point size, log2FC cut-off and FDR cut-off are selectable and colors of DEGs are customizable (*Fig 14*).

In comparison with the MA plot, the Volcano plot separates the up and down regulated genes with respect to the log2FC cut off. The result table that is similar to the one produced for the MA plot is made available to the user, according to the selected FDR cutoff, as well as the down- and up-regulated genes and their associated data, respectively.

These two plots are only available for comparative analysis between two conditions ,as they are built using log2FC and BaseMean. If the DE analysis includes three conditions or more, the MA and Volcano plot will not be available and an error message will occur.

In order to further visualize the results, I added a heatmap that can also be used for clustering and further analysis time-series data (discussed below). And finally, another PCA, realized exactly the same way as the one using raw data, but this time using the normalized data to again check the groups with the possibility to restrict clustered genes by choosing an FDR cutoff. This PCA is available in 2D and 3D. The user can rotate the 3D PCA and a static image can be downloaded.

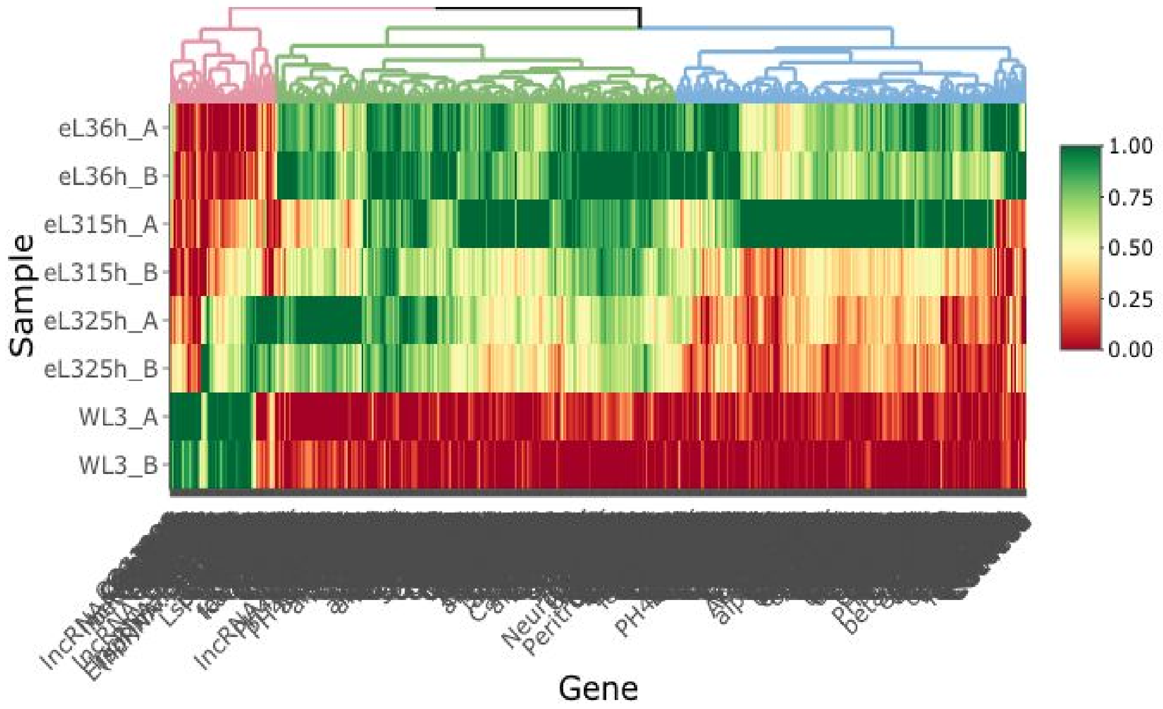


Fig 15 : Illustration of Heatmaply heatmap with its clusters colored on the dendrogram.

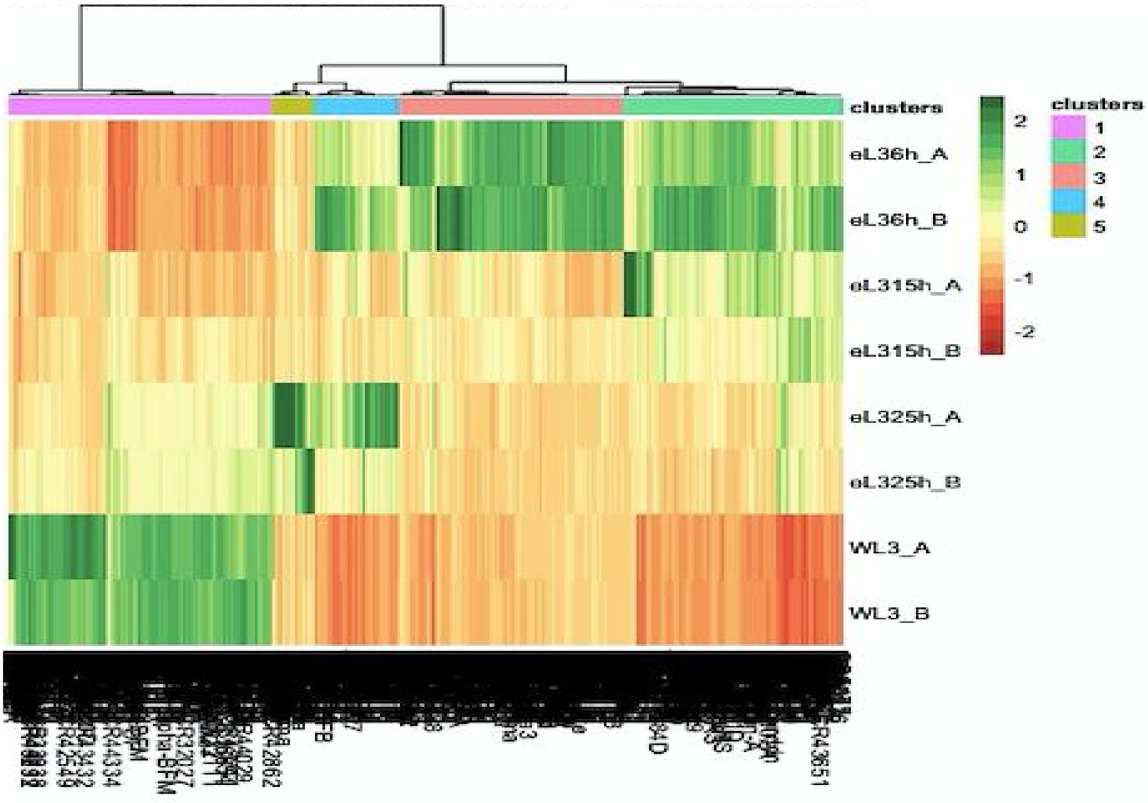


Fig 16 : Illustration of pheatmap heatmap using clustering. Clusters are colored and named.

4.1.2.4 - Clustering of time-series data

As our collaborators performed a time-series study, I wanted to give them the possibility to perform analysis over the time-points. Profile-based clustering methods like mfuzz [25] are difficult to use for non-specialists, as they require manyfold tests for selecting the correct cluster number, membership values that should be chosen, etc. Thus, I decided to use a simple heatmap-based clustering to build clusters from time-series data. To use a heatmap-based clustering is a simple, but potentially efficient method that can be easily applied by biologists. The advantage of a heatmap-based clustering is also that the number of clusters results from the clustering process of the data itself and the user does not have to guess a number of different clusters.

The `Heatmaply` function from the eponym R-package was again an obvious choice.
It offers the possibility to color the dendrogram with as many colors (i.e. clusters) as wanted with the `k_col` argument, resulting in more or less clusters. To create the table of associated genes with a cluster, I used the `h_clust` function with the same distance and agglomeration measures as for the heatmap, as heatmaply uses `h_clust` to make its dendrogram. Then I used the `cutree` function to cut the dendrogram in the wanted number of clusters and finally added it to the `k_col` argument of `heatmaply`. This allows a dendrogram colored with respect to the clusters, as well as the associated gene tables. It also allows the user to choose an appropriate number of clusters depending on the resulting heatmap and the given data.

The major problem of clustering with `heatmaply`, is that the colors are not associated with the clusters given in the table, as these are numbered. It thus forces the user to check at least one gene of each cluster to associate it with the numbered clusters in the table. This might work with a few clusters, but is not very user friendly and not feasible with more than 10 clusters (*Fig 15*).

Thus we are not satisfied with the clustering function of the `heatmaply` package. I therefore implemented the `pheatmap` heatmap function from the eponym R-package. While this function is not working with `plotly`, it allows to cluster genes and make the clusters clearly visible and identifiable in three steps, them being named and easily associable in the resulting table (*Fig 16*).

4.1.2.5 - GO enrichment & ID conversion

The final functionality we decided to add to the app is functional enrichment. I wanted to perform the enrichment with a gene list pasted in the app. I tested two different methods for enrichment, the `enrichGO` and the `enrichR` function.
At first, I tested the `enrichGO` function from the `clusterProfiler` R-package [26].

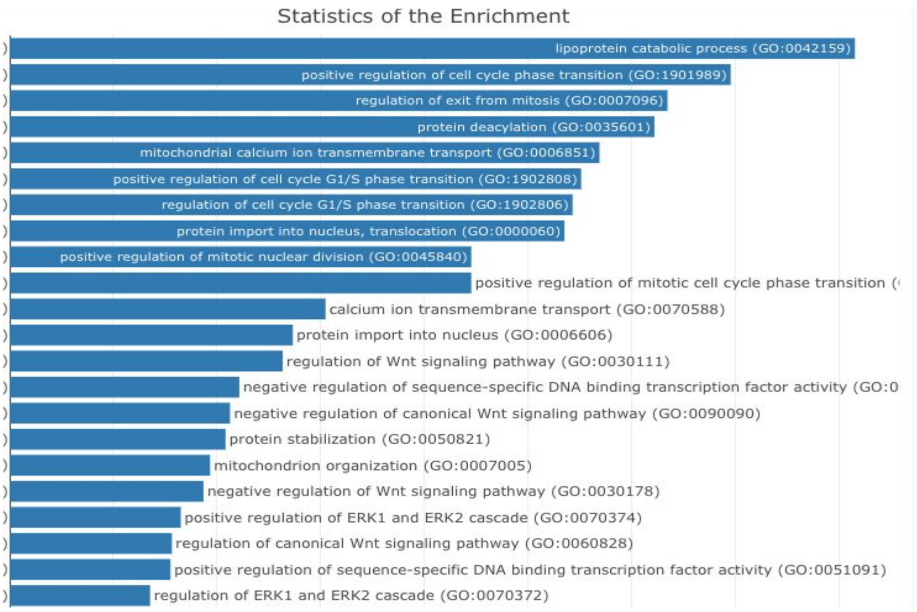


Fig 17 : Illustration of the GO enrichment using enrichR. Visualization available in bar plot with respect to -log10(p-value).

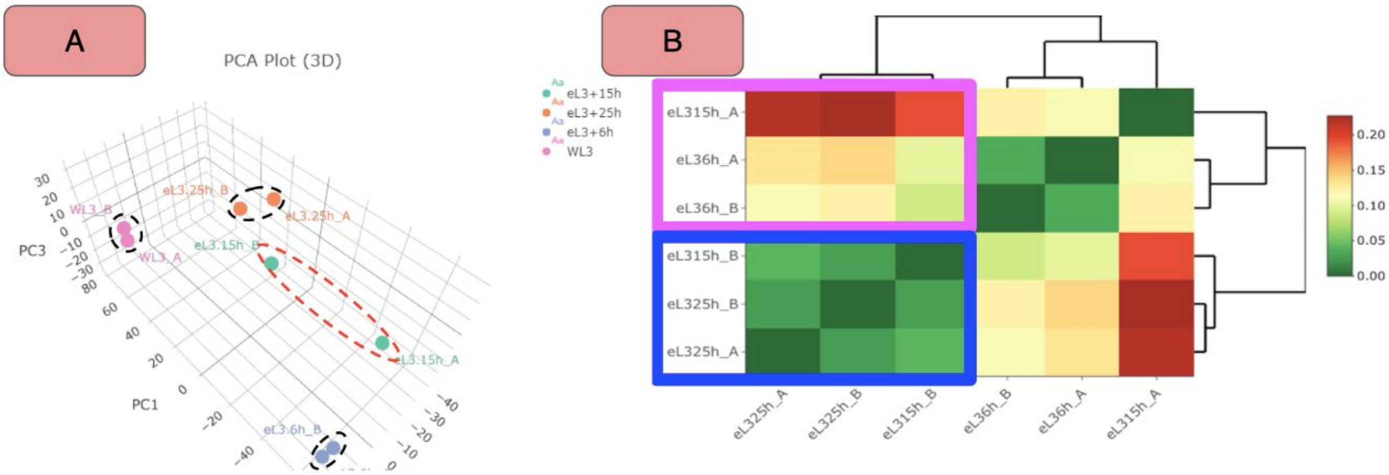


Fig 18 : Validation of sequencing libraries. Two replicates were sequenced per time point. Every group is well correlated in the 3D PCA except the eL3+15h group (A). Hierarchical clustering of groups to establish which of the eL3+15h replicates is problematic. The pink square shows that replicate A associates more strongly with the eL3+6h group and the blue square shows that replicate B associates with the eL3+25h group (B).

enrichGO offers to choose the ontology to perform enrichment with. One can choose to see all resulting enriched ontologies at once or make a selection by precising a p-value and a possible adjustment of this p-value by multiple testing, which is useful to remove false positive results [27].

What was not very efficient and practical was that enrichGO required to install the individual annotation for each wanted organism, so I had to make a choice of which organisms to add. For the moment, I have added *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans* and *E. coli*.
The second package I tested was enrichR [17] , which exists as a web-server, but also as an R-package. The enrichR package only needs two code lines: one for geneset declaration; and the other one to perform the enrichment. The organism does not need to be declared. It works with any organism. A single ontology has to be chosen and the results are stored in a table. I chose a barplot to visualize the GO enrichment, showing GO terms and associated -log10(p-value), as well as a pie chart to offer another visualization for data mining (*Fig 17 A&B*).

The reason I added an ID conversion tool to the app was that biologists prefer to work with gene symbols, rather than identifiers. I provide a conversion tool using the bitr function from the clusterProfiler R-package. This tool converts IDs between EntrezIDs, EnsemblIDs, and gene symbols. However, as it is part of clusterProfiler, the annotation R-package of the organism needs to be specified. For the moment, I keep the five organisms described earlier (human, mouse, *Drosophila*, *C. elegans* and *E. coli*). This can however be easily extended.

4.1.3 - Biological Results

In this paragraph, I will demonstrate the functionality of the RNApp using the biological data produced in the Maurange team. First, it validates the sequencing libraries and the quality of the replicates per time point. Every replicate should correlate with its respective group. As is shown in Fig 18 A, there is a problem with the third time-point. Indeed, the two replicates of eL3+15h time point are separated in the PCA (*Fig 18 A*).

The two replicates are indeed different and the Maurange team found in DEGs in a comparison between the first group (eL3+6h) and the second (eL3+15h), genes corresponding to the larva cuticle in the eL3+15h group, especially present in the replicate A. A contamination of the larva cuticle likely has occurred in the first replicate, while the second replicate is closer to the third group (eL3+25h).

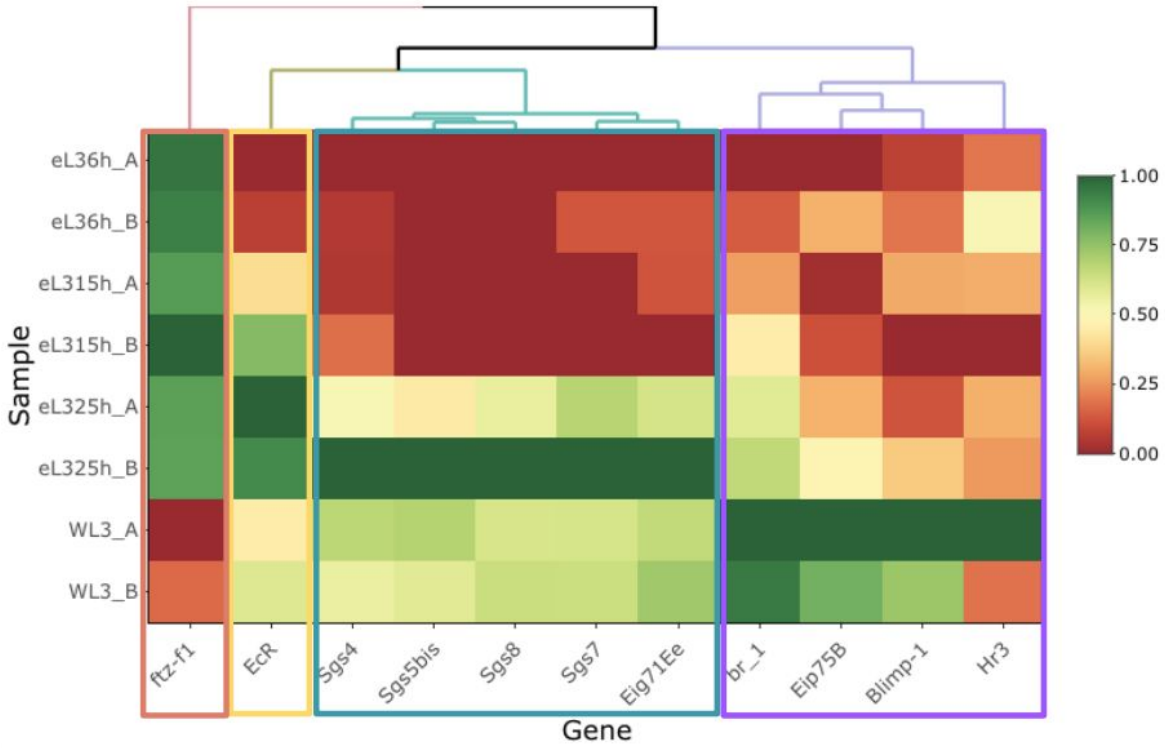


Fig 19 : Heatmap of Ecdysone induced genes, showing that transcriptomic variation occurs between the eL3+35h and WL3 groups. Focusing on the red cluster of ftz-f1,a transcription factor of the metamorphosis and the violet cluster of Blimp-1, a transcription factor induced by Ecdyson and having a role in development of organs. In the Sgs_ cluster are majorly glue genes.

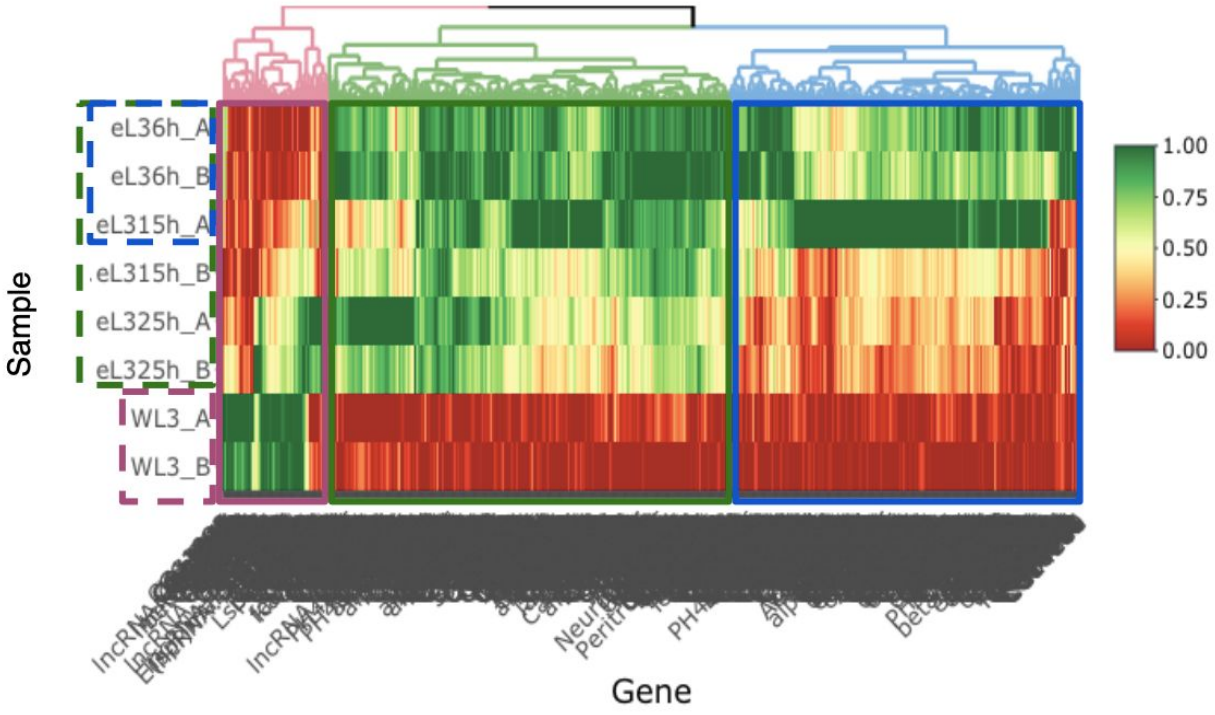


Fig 20 : A heatmap of all 2156 DEGs with a FDR cutoff of 1e-5 from the 9081 total genes analyzed after removing genes with expression < 100 from the 15349 input genes. The blue cluster corresponds to the early L3 stage, the Chinmo stage. The green cluster is the transition cluster where we can see genes slowly activated and finally the pink cluster is the differentiation cluster, containing the Broad-Z1.

The Maurange team decided to continue the analysis for the moment with the two replicates before finding a better solution as it needs 2 replicates for RNApp.

In a global analysis of Ecdysone induced genes using heatmap clustering, we could observe that the real variation in transcriptomic programs occured between the two last groups (eL3+25h and WL3) (*Fig 19*). The transcription factor **ftz-f1** bridges early and late gene expression during the process of metamorphosis and is essential to the development in *D. melanogaster*. **Blimp-1** is a ZBTB transcriptional repressor of ftz-f1 that is induced by Ecdysone and plays important roles for metamorphosis and more precisely times Ecdysone development pathways including regulation of ftz-f1 timing [28]. In the Sgs_ cluster, the genes included are majorly glue genes. Glue is a mixture of unrelated proteins allowing insects to adhere to wood, leaves, and other surfaces but also is required in pupa adhesion [29].
As we can see in the Figure 19, ftz-f1 only acts during the metamorphosis and during the development and Blimp-1, only in WL3 at the end of the metamorphosis, which correlates with their role previously described.

Regarding all DEGs, we were expecting 3 clusters: one for the early L3 stage - the Chinmo cluster; one for the mid stage, which is the transition; and the last cluster for the later L3 stage. We could identify these three clusters with a small last one. This time point is very late, and the resulting cluster - the Broad-Z1 cluster, should contain only differentiation related genes (*Fig 20*).

4.2 - ATAC-seq analysis

ATAC-seq libraries were sent to sequencing to discover changes of the chromatin landscape over time. To analyze the data, I built a command-line based pipeline to analyse the data starting from raw fastq files to produce the final peak data (*Fig 21*). The shell-based pipeline can be executed in one go, but as steps can be optional, it can be used command by command and it can be modified.

I began the classic preprocessing with a FastQC quality control of the reads and a trimming of the adapters and low quality reads. This step is potentially required as paired-end sequencing may lead to adapter-only fragments, because they are smaller than 2x50 bp. Adapter trimming is optional and depends on the results of the first FastQC analysis. Indeed, if the user suspects contamination with adapters based on the size distribution of the libraries (e.g. many short, identical reads), adapters should be removed.

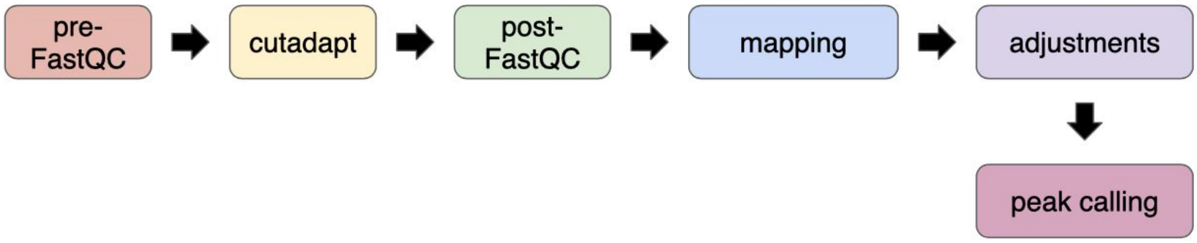


Fig 21 : ATAC-seq analysis pipeline. For mapping Bowtie2 is used; MACS2 is used for peak calling.

For trimming, cutadapt offers a lot of options regarding the 5’ or/and 3’ adapter sequences and is the most popular in the category. This step is followed by a second FastQC to check the results of trimming.

The two most popular softwares for mapping ATAC-seq data are Burrows-Wheeler Alignment (BWA) [30] and Bowtie2 [14]. As Bowtie2 is one of the fastest short read mapping programs, I chose it. Before mapping, creating the genome index is essential. The `bowtie2-build` function is dedicated to this step where Bowtie2 indices are made from a fasta genome file of the organism of interest. The `bowtie2` function performs the mapping and I chose the `--very-sensitive` option to have more chance to hit the best matches. With this option, the following parameters are chosen:
-D - a parameter for choosing the attempts done for seed extension to find a better match before moving to the next one has to be chosen, I chose 20, which is the default.
- R 3, 3 maximum reseeding attempts, meaning choosing a new set of reads of the same length to search for more alignments.
-N 0, 0 mismatches allowed in a seed alignment in multi-seed alignments.
-L 20, length of the substring to align during multi-seed alignments.
-i S,1,0.50 sets the interval function f where x is the read length to

$$f(x) \ = \ 1 + 0.5 * \sqrt{x}$$

The output of bowtie2 mapping is a sam file. To make adjustments such as removing the mitochondrial chromosome as it doesn’t contain peaks, I needed to convert it to a bam file. I used `samtools view -h` to make the conversion and then used `grep -v` to grab the chrM chromosome and `samtools view -b -q 10` to remove non unique alignements.

Peak calling is a computational method used in ChIP-seq analysis but also with ATAC-seq data to identify enriched regions in aligned reads after mapping to the genome. These regions are indicative of protein binding and are called “peaks”. To perform peak calling using MACS2, a bed file is needed. When MACS2 shifts the reads to the center of a binding site, it will only consider one of the read pairs. I reused `samtools view -h` to convert the filtered bam file to a sam file again; and then the `SAMtoBED` function from Harvard [31] to convert to a bed file.

Peak calling is done with the `MACS2 callpeak` function. To avoid over-calling of a peak, MACS2 also provides options to deal with duplicate tags at the exact same location. After shifting every tag by d/2, where d is the distance between two peaks summit, MACS2 slides across the genome to find candidate peaks in a 2d window and the tag distribution can be modeled as a Poisson distribution.

$$P_{\lambda}(X = k) = \frac{\lambda^k}{k! * e^{-k}}$$

where λ is the expected number of reads in the current window.

Each peak is an independent test and a FDR is calculated for each peak. For multi comparison with multi testing, the p-value is adjusted using the Benjamin Hochberg correction.
The most relevant output files of MACS2 are 1)_ a narrowPeaks file, which reports peaks signal enrichments; and 2)_ a bed file containing the peak summit locations that can be uploaded in IGV to seek motifs of TF binding sites.

5 - DISCUSSION

I built a complete RNA-seq analysis process including preprocessing, mapping and read counting; as well as a user-friendly and interactive R-shiny app for RNA-seq differential expression analysis and clustering, which includes ID-conversion, as well as gene ontology enrichment. The ID-conversion and enrichment function I have also built as a small R-shiny app named R-enrichTool, which is available on my GitHub @margauxhaering.
Making such an app, there are always choices for specific tools that have to be made. Some of the tools used could be improved or exchanged, depending on the wishes of the users. The clustering function implemented in RNApp is a very simple one. However, it allows biologists to get an easy overview of the gene clusters behaving similarly e.g. over a time-series experiment. Other than mfuzz or other clustering methods, the user does not need to perform additional steps to test the validity of the resulting profile clusters. Then, choosing to also implement the `pheatmap` function, even though it produces less nice figures, allowed our collaborators to easily identify which genes belong to which cluster. About the ID conversion tool, a better tool probably exists for which the organisms to be included do not need to be specified.

Focusing on an improvement of the ATAC-seq pipeline is a next goal. Other tools to perform a peak calling are available and especially designed for ATAC-seq: Genrich [32] (not yet published) and HMMRATAC [33]. HMMRATAC uses Hidden Markov Models (HMM) and is the first dedicated ATAC-seq analysis tool published to date.
HMMRATAC segments the genome into three states, open chromatin regions with high signals, moderated signals in the nucleosome regions and low signals in the background regions. It uses HMMs to predict open chromatin regions, decomposing the dataset into layers of coverage signals and relating the layers to each other.

I want to build a specific ATAC-seq data analysis pipeline using HMMRATAC. I also want to generate a visualization part with figures to illustrate HMMRATAC results. I will probably realise this also in a R-Shiny app to advance ATAC-seq analysis further in the project.

The Maurange team is planning on doing a TaDa experiment to get an even more complete picture of the transition. DNase Adenine Methylase Identification (DamID) [34] generates genome-wide maps of chromatin protein binding by fusing a Dam molecule to the protein of interest. Targeted DamID (TaDa) is a DamID profiling which identifies transcribed genes using the phenomenon of ribosome reinitiation to express Dam-fusion proteins [35] . Performing a TaDa analysis would allow us to discover direct Chinmo and Broad-Z1 target genes.

Integrated with ATAC-seq and RNA-seq results, it should shed light on the effect of an open or a closed chromatin state generated by a certain gene on Chinmo and Broad-Z1 target gene regulation. In the near future, building a TaDa analysis pipeline will therefore be necessary.

6 - CONCLUSION

For my thesis, I created a set of analysis pipelines to analyze RNA-seq and ATAC-seq data. The RNA-seq analysis pipeline consists of a preprocessing pipeline including a quality control, a mapping and feature counting; and an R-Shiny app, RNApp, which performs RNA-seq statistical and differential expression analysis, clustering of samples using hierarchical clustering, enrichment analysis as well as a rich set of data visualization options. It includes the TCC R-package providing several methods of normalization and DEG identification and offers downloadable figures and tables.

RNApp is very user friendly, doesn’t require coding from the user, allowing data analysis from raw read count to GO enrichment and ID conversion, if necessary.

RNApp allowed the Maurange team to be independent in data analysis and to identify clusters of genes of interest within the transition between self-renewal and differentiation in *D.melanogaster*.

As ATAC-seq data are coming, I built a modular ATAC-seq pipeline, with selectable steps. It starts from raw fastq files to peak calling, where the output is a bed file that can be visualized later in IGV.

My internship allowed me to put in application knowledge I acquired in my Master’s degree but also pushed me to investigate and find accurate solutions to biological data analysis problems.

REFERENCES

[1] : Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484. PMID: 19015660; PMCID: PMC2949280.

[2] : Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213-8. doi: 10.1038/nmeth.2688. Epub 2013 Oct 6. PMID: 24097267; PMCID: PMC3959825.

[3] : Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: an R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics. 2013 Jul 9;14:219. doi: 10.1186/1471-2105-14-219. PMID: 23837715; PMCID: PMC3716788.

[4] : Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24-6. doi: 10.1038/nbt.1754. PMID: 21221095; PMCID: PMC3346182.

[5] : Habermann B,Marchiano F, Meiler A. AnnoMiner. IBDM, 2019.

[6] : Yamanaka N, Rewitz KF, O'Connor MB. Ecdysone control of developmental transitions: lessons from Drosophila research. Annu Rev Entomol. 2013;58:497-516. doi: 10.1146/annurev-ento-120811-153608. Epub 2012 Oct 15. PMID: 23072462; PMCID: PMC4060523.

[7] : Narbonne-Reveau K, Maurange C. Developmental regulation of regenerative potential in Drosophila by ecdysone through a bistable loop of ZBTB transcription factors. PLoS Biol. 2019 Feb 11;17(2):e3000149. doi: 10.1371/journal.pbio.3000149. PMID: 30742616; PMCID: PMC6386533.

[8] : Khan SJ, Abidi SNF, Skinner A, Tian Y, Smith-Bolton RK. The Drosophila Duox maturation factor is a key component of a positive feedback loop that sustains regeneration signaling. PLoS Genet. 2017 Jul 28;13(7):e1006937. doi: 10.1371/journal.pgen.1006937. PMID: 28753614; PMCID: PMC5550008.

[9] : Harris RE, Setiawan L, Saul J, Hariharan IK. Localized epigenetic silencing of a damage-activated WNT enhancer limits regeneration in mature Drosophila imaginal discs. Elife. 2016 Feb 3;5:e11588. doi: 10.7554/eLife.11588. PMID: 26840050; PMCID: PMC4786413.

[10] : Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[11] : MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. Date accessed: 30 apr. 2020. doi:https://doi.org/10.14806/ej.17.1.200.

[12] : Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.

[13] : Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

[14] : Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.

[15] : Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. doi: 10.1186/gb-2008-9-9-r137. Epub 2008 Sep 17. PMID: 18798982; PMCID: PMC2592715.

[16] : Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012 Sep;7(9):1728-40. doi: 10.1038/nprot.2012.101. Epub 2012 Aug 30. PMID: 22936215; PMCID: PMC3868217.

[17] : Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(W1):W90-7. doi: 10.1093/nar/gkw377. Epub 2016 May 3. PMID: 27141961; PMCID: PMC4987924.

[18] : Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. doi: 10.1186/gb-2010-11-3-r25. Epub 2010 Mar 2. PMID: 20196867; PMCID: PMC2864565.

[19] : Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

[20] : McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012 May;40(10):4288-97. doi: 10.1093/nar/gks042. Epub 2012 Jan 28. PMID: 22287627; PMCID: PMC3378882.

[21] : Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug 10;11:422. doi: 10.1186/1471-2105-11-422. PMID: 20698981; PMCID: PMC2928208.

[22] : Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. Algorithms Mol Biol. 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5. PMID: 22475125; PMCID: PMC3341196.

[23] : Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

[24] : Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 2016 Feb 25;17:103. doi: 10.1186/s12859-016-0956-2. PMID: 26911985; PMCID: PMC4766705.

[25] : Kumar L, E Futschik M. Mfuzz: a software package for soft clustering of microarray data. Bioinformation. 2007 May 20;2(1):5-7. doi: 10.6026/97320630002005. PMID: 18084642; PMCID: PMC2139991.

[26] : Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28. PMID: 22455463; PMCID: PMC3339379.

[27] : Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values? Cell J. 2019 Jan;20(4):604-607. doi: 10.22074/cellj.2019.5992. Epub 2018 Aug 1. PMID: 30124010; PMCID: PMC6099145.

[28] : Agawa Y, Sarhan M, Kageyama Y, Akagi K, Takai M, Hashiyama K, Wada T, Handa H, Iwamatsu A, Hirose S, Ueda H. Drosophila Blimp-1 is a transient transcriptional repressor that controls timing of the ecdysone-induced developmental pathway. Mol Cell Biol. 2007 Dec;27(24):8739-47. doi: 10.1128/MCB.01304-07. Epub 2007 Oct 8. PMID: 17923694; PMCID: PMC2169387.

[29] : Borne F, Kovalev A, Gorb S, Courtier-Orgogozo V. The glue produced by *Drosophila melanogaster* for pupa adhesion is universal. J Exp Biol. 2020 Apr 23;223(Pt 8):jeb220608. doi: 10.1242/jeb.220608. PMID: 32165432.

[30] : Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

[31] : Gaspar J. M. SAMtoBED, 2017. Available at : https://github.com/jsh58/harvard/blob/master/SAMtoBED.py

[32] : Gaspar J. M. Genrich, 2018. Available at : https://github.com/jsh58/Genrich

[33] ：Tarbell ED, Liu T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. Nucleic Acids Res. 2019 Sep 19;47(16):e91. doi: 10.1093/nar/gkz533. PMID: 31199868; PMCID: PMC6895260.

[34] : Maksimov DA, Laktionov PP, Belyakin SN. Data analysis algorithm for DamID-seq profiling of chromatin proteins in Drosophila melanogaster. Chromosome Res. 2016 Dec;24(4):481-494. doi: 10.1007/s10577-016-9538-4. Epub 2016 Oct 21. PMID: 27766446.

[35] : Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, Marshall OJ, Brand AH. Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. Dev Cell. 2013 Jul 15;26(1):101-12. doi: 10.1016/j.devcel.2013.05.020. Epub 2013 Jun 20. PMID: 23792147; PMCID: PMC3714590.

ABSTRACT

In my internship, I worked on the development of a complete and flexible RNA-seq, as well as ATAC-seq pipeline. Being embedded between a biology and bioinformatics team, my task was to develop first a modular pipeline for RNA-seq data processing, from quality control to read mapping and feature counting but also a user friendly interactive R-shiny app for RNA-seq data analysis. This app, which I called RNApp, takes raw count files and performs a differential expression analysis, as well as a GO enrichment. It offers different tools for normalisation and differential expressed genes identification, clustering of genes over different time-points or conditions and a rich set of interactive visualizations are available for data exploration. As this project was done in the course of a collaboration with a biological team, I demonstrate how the app can be used with *Drosophila melanogaster* developmental study data. As next to RNA-seq data, ATAC-seq data will also need to be analyzed, I developed in a second part of my project a modular ATAC-seq pipeline starting from quality control of reads to read mapping and peak calling.