Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status

Student: Víctor Sánchez Gaya
Master's degree in Bioinformatics and Biostatistics
Area 37: Omics data analysis and integration

UOC consultant: Sonia Tarazona Campos
Responsible of the subject: Jose Antonio Morán Moreno

5th June 2018

FINAL PROJECT WORKSHEET

Project Title:	*Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status*
Author's name:	*Víctor Sánchez Gaya*
Tutor's name:	*Sonia Tarazona Campos*
PRA's name:	Jose Antonio Morán Moreno
Deadline:	06/2018
Degree's name:	*Master's degree in Bioinformatics and Biostatistics*
Final Project Area:	*37, Omics data analysis and integration*
Language:	*English*
Keywords:	*Omics integration, ChIP-Seq, RNA-Seq*

Abstract (in English, 250 words or less):

The Yeast Metabolic Cycle (YMC) has become a model to study how changes in the metabolic landscape can affect the chromatin status to regulate gene expression. Previous studies have shown how the YMC is divided in three main phases and have described the effect of certain histone modifications in gene regulation. This project constitutes a novel application of different statistical methodologies for the integration of chromatin status data (ChIP-Seq) and gene expression data (RNA-Seq) to better understand the YMC.

The usage of regression models (N-PLS and MORE methodologies) for the omics integration allowed us for assessing the relevance of histone modifications and transcription factors on the regulation of gene expression changes in the YMC. H3K18ac and H3K9ac turned out to be the most important of the studied histone modifications, whereas YLR403W, YPL254W, YOR363C, YGL209W and YDR451C emerged as the most relevant transcription factors. A significant association of co-regulation of gene expression was found between H3K18ac and the transcription factors YPL254W (involved in the formation of the SAGA complex) and YLR403W (involved in the process of histones exchange), which evinced the crucial role of the acetylation levels to regulate gene expression in the YMC through a coordinated action of transcription factors and histone modification levels.

Thus, in this study, new connections were established between metabolome, chromatin status and gene expression, as well as the basis to identify potential regulators of hidden regulatory mechanisms that connect histone modifications and gene expression changes.

Resum:

El Cicle Metabòlic del Llevat (YMC) s'ha convertit en un model d'estudi per analitzar la influència dels canvis metabòlics sobre l'estat de la cromatina per a la regulació de l'expressió gènica. Estudis previs han demostrat que el YMC es divideix en tres fases principals i han descrit l'efecte de certes modificacions d'histones en la regulació gènica. Aquest projecte constitueix una aplicació innovadora de diferents metodologies estadístiques per a la integració de dades d'estat de la cromatina (ChIP-Seq) i d'expressió gènica (RNA-Seq).

L'ús de models de regressió (metodologies N-PLS i MORE) per a la integració d'òmiques ens ha permés avaluar la rellevància de les modificacions d'histones i els factors de transcripció en la regulació dels canvis d'expressió gènica en el YMC. H3K18ac i H3K9ac van resultar ser les més importants de les modificacions estudiades, mentre que YLR403W, YPL254W, YOR363C, YGL209W i YDR451C van ser els factors de transcripció més rellevants. Es va trobar una associació significativa de co-regulació de l'expressió gènica entre H3K18ac i els factors de transcripció YPL254W (implicat en la formació del complex SAGA) i YLR403W (involucrat en el procés d'intercanvi d'histones), que evidenciava el paper crucial dels nivells d'acetilació a l'hora de regular l'expressió gènica al YMC mitjançant l'acció coordinada dels factors de transcripció i els nivells de les modificacions d'histones.

D'aquesta manera, en aquest estudi s'estableixen noves connexions entre el metaboloma, l'estat de la cromatina i l'expressió gènica, així com la base per identificar els possibles reguladors dels mecanismes de regulació que connecten les modificacions d'histones amb els canvis d'expressió gènica.

Acknowledgments

És un gran gust poder viure rodejat d'amics, i sempre és bon moment per a donar-lis les gràcies.

Ha sigut tot un plaer poder estar de nou al grup d'Ana per a desenvolupar aquest projecte, agraisc infinitament l'oportunitat d'haver format part d'ell i el tracte rebut, ja que no sols m'ha permès seguir formant-me acadèmicament d'una forma que no haguera pogut imaginar, sinó conèixer gent que espere seguir tenint a prop per molts anys més, com per exemple a Manu, per a què seguisca portant-me al tren, i a Carlos, m'encanten les teues empanadilles de pisto, no deixes de fer-les per favor. M'agradaria remarcar especialment l'esforç d'aquelles persones sense les quals aquest treball haguera sigut inviable. En primer lloc a Salva, el primer oficial al camp de batalla, gràcies amic (encara què de vegades et mataria). En segon lloc, i com no, a Sonia, la persona que probablement més haja contribuït a la meua formació acadèmica, estic segur de què no haguera pogut trobar a una millor persona de la que poder seguir aprenent, gràcies per la teua paciència i bon rotllo amiga (encara què de vegades em mataries, i ho saps jaja).

Als meus pares, per no només ser pares, sinó també amics. I como no a la meua compitruein Maria, la meua parella indispensable d'aventures.

Index

Figures Index

Tables Index

1. Introduction

1.1 Context and justification of the project

Recent advances in omics technologies offer new frameworks to approach biological questions. Today each type of biomolecular entity can be studied by a different omic method, such as genomics, epigenomics, transcriptomics or metabolomics. Although the individual study of each omics type provides interesting insights, it is limited to address the regulation of the whole system where multiple layers constantly interact. The joint analysis of different –omics has gained ground in recent years, but the integration of different omics into a single statistical model is still a relevant challenge in the field of bioinformatics.

This master thesis constitutes a novel application of different statistical methodologies for the integration of chromatin status data (ChIP-Seq) and gene expression (RNA-Seq) focused on the study of the Yeast Metabolic Cycle (YMC). YMC is given under continuous glucose-limited conditions, and it is defined by a robust periodic change of gene expression, which drives the cell towards a cyclic response to respiratory oscillations. The YMC has been linked to cell cycle or circadian rhythms, and recently, chromatin modifications have been suggested to regulate the genes whose expression defines YMC (Kuang, et al., 2014). Although some studies have shown a correlation between chromatin modifications and gene expression within the YMC, the exact regulatory and functional mechanisms are still not well understood. The study of such mechanisms can shed some light to the underlying processes that allow organisms to respond to different metabolic conditions. The aim of this project is to understand how the chromatin modifications control gene expression, and to reproduce the regulatory networks that contribute to the cyclic cell response during the YMC.

The data used in this master thesis were published in (Kuang, et al., 2014). In that project, the putative effect of different histone modifications over the major oscillations of gene expression found in the YMC was analysed. This master thesis was developed in the Genomics of Gene Expression group at Centro de Investigación Príncipe Felipe (València). Since the group was interested in getting more insights about the YMC, we took the data in this paper as a starting point and tried to improve and complete the study that had already been done, which had not used integrative strategies to combine gene expression and chromatin status information.

According to (Kuang, et al., 2014) the YMC can be divided into three phases on the basis of the gene expression profiles: oxidative (OX), reductive building (RB) and reductive charging (RC). Growth genes, such as ribosomal and amino acid–biosynthesis genes are activated in OX phase; mitochondria and cell-cycle genes are expressed in RB; and genes responding to starvation, stress and survival are elevated in RC (Kuang, et al., 2014). Along this project, the conjecture regarding the existence of these 3 phases will be assumed.

1.2 Objectives

The main objective of the project is the development of an appropriate omics integration analysis pipeline to elucidate the role of chromatin status on the regulation of gene expression in the Yeast Metabolic Cycle.

In order to achieve this objective, the following secondary objectives will also have to be accomplished:

- Application of an appropiate pre-processing and preparation strategy to the omics data (RNA-seq and histone modification ChIP-seq), which is key for obtaining meaningful results from the integrative analysis

- Selection of the differentially expressed genes by using an appropiate tool that deals with time-series data, and clustering the genes according to their temporal profiles.

- Usage of suitable regression methods to model the relationship between gene expression and histone modifications to gain insights in global and specific gene expression regulation across the Yeast Metabolic Cycle.

- Construction of a network to depict and summarize the relevance of the regulators found in the previous step.

- Biological interpretation of the results to compare with previously reported knowledge and highlight new findings.

1.3 Approach and methodology

As indicated in Section 1.1, this project aims to study the role of the chromatin status (ChIP-Seq data) over the regulation of the gene expression (RNA-Seq data) in the YMC.

This analysis can be performed in two different ways. On the one hand, all the analyses of RNA-Seq and ChIP-Seq data can be carried on independently or in parallel, and afterwards combine the results from both omics and interpret them. This type of analysis is known as conceptual or sequential integration (Cavill, et al., 2016). On the other hand, a simultaneous analysis of both omics could be applied, through the usage of statistical methodologies which can integrate in the same model the different omics data. The benefits of this statistical integration approach above the conceptual integration is that new relationships among the studied omics can be detected, that cannot be found in another way. Hence, taking into account the complexity of the living organisms, which can be seen as a complex network of biological modules or layers which interact in a coordinated way, the independent analysis of each omics data is not the best choice.

The simultaneous modelling of different omics to reveal the relationships among them is still a field under development, due to the difficulties raised when dealing with variables measured with different technologies, measurement errors, variances… Hence this thesis has been a challenging project which expects to contribute to the improvement of this field.

The statistical models used for the integration of the data were N-PLS and MORE. A detailed description of the methodology applied in this project is given in Section 2 (Materials and methods), and an overview of the followed pipeline is depicted in Figure 7.

1.4 Work plan

In this section, a time planning of the tasks is given, regarding the different PECs: PEC0, PEC1, PEC2, PEC3, PEC4 y PEC5.

PEC0 (from 21/02/18 to 05/03/18)

Approximately 2 weeks.

Table 1. Gantt chart of the tasks performed during the PEC0. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	Week 2
Discussion of the topic of the project	🟩	
Data obtaining from a public database		🟩

PEC1 (from 06/03/18 to 19/03/18)

Approximately 3 weeks.

Table 2. Gantt chart of the tasks performed during the PEC1. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	Week 2	Week 3
ChIP-Seq raw data pre-processing: quality control, trimming, mapping, reads genome coverage, normalization…	🟩	🟩	🟩
Chip-Seq pre-processed data exploration			🟩
RNA-Seq data exploration			🟩
Multiway techniques study (reading bibliography)		🟩	🟩
N-PLS (reading bibliography)		🟩	🟩
Regression modelling techniques study (reading bibliography)		🟩	🟩
Workplan document elaboration			🟩
Workplan document delivery			🟩

PEC2 (from 20/03/18 to 23/04/18)

Approximately 4 weeks.

In the PEC1 submitted document, it was explained that the MORE tool was going to be used to model gene expression for each differentially expressed gene, considering the histone modifications as predictors, to find out which histone modifications were more relevant for the regulation of each specific gene, but also globally when analysing simultaneously all the collected data with the N-PLS model. The relevant addition to notice is that along this PEC2 it was decided to analyse, not only the influence of histone modifications on gene expression, but also the regulation of transcription factors, in order to study their role in the YMC, and their effect in combination with the chromatin status data. The application of MORE with transcription factors data was done in the PEC3.

Table 3. Gantt chart of the tasks performed during the PEC2. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	Week 2	Week 3	Week 4
RNA-Seq modelling: differential expression and clustering analysis	🟩			
Batch effect study and correction		🟩		
Omics integration		🟩	🟩	🟩
N-PLS application		🟩	🟩	
MORE application with Histone				🟩

Modifications data				🟩
Master thesis manuscript writing		🟩	🟩	🟩

PEC3 (from 24/04/18 to 21/05/18)

Approximately 4 weeks.

After the application of MORE with transcription factors, it was also decided to create a regulatory network to represent graphically the interaction of the most relevant transcription factors and the different studied histone modifications, to better understand the regulation of gene expression during the YMC. Thus, the creation of the regulatory network constituted an extra-step not considered at the beginning of the project.

It was also analysed the significance of the joint action between each transcription factor and each histone modification. The idea was to study whether a specific transcription factor was significantly regulating the same group of genes that was being regulated by a specific histone modification. This step was another task not considered initially, since the study of transcription factors was not planned at the beginning of the project.

Table 4. Gantt chart of the tasks performed during the PEC3. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	Week 2	Week 3	Week 4
MORE application with Transcription Factors data	🟩			
Selection of the most relevant Transcription Factors	🟩			
Significant joint action analysis: TF-histone modification	🟩			
Gene set enrichments analysis based On the MORE results		🟩		
Regulatory network : TF-Hist Mod			🟩	
Biological results profound analysis and interpretation			🟩	🟩
Master thesis manuscript writing	🟩	🟩	🟩	🟩

PEC4 (from 22/05/18 to 05/06/18)

Two weeks.

Table 5. Gantt chart of the tasks performed during the PEC4. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	Week 2
Master thesis manuscript writing	🟩	
Improving figures quality	🟩	🟩
Reviewing citation and manuscript format		🟩

Master thesis manuscript delivery		🟩

PEC5 (from 06/06/18 to 25/06/18)

Approximately 3 weeks.

Table 6. Gantt chart of the tasks performed during the PEC5. The green cells represent in which week the task described at the left side of the row was carried out.

	Week 1	**Week 2**	**Week 3**
Preparation of the slides for the thesis public defense	🟩		
Preparation of the speech for the thesis public defense	🟩		
Public defense of the thesis when indicated by the UOC		🟩	🟩

1.5 Brief summary of the products obtained

The different documents produced and submitted for the evaluation of this master thesis are:

- The **PEC1**, **PEC2** and **PEC3** delivered files, where the work plan of the time period comprehended between 05/03/2018 and 21/05/2018 is shown. That information is also described in Section 1.4 of this document.

- The **master thesis manuscript**, which corresponds to this document. It contains a detailed description of all the methodology applied in the project and the results obtained.

- A **slides presentation** will be prepared for the public defense of the project the week after submitting the master thesis manuscript.

In addition:

- In parallel to this master thesis, a **scientific publication** based on this work has been also prepared that will be submitted to **Frontiers in Genetics** journal.

- All the **scripts** produced will be uploaded to a public **BitBucket** folder, for the benefit of the researchers interested in developing similar analyses.

1.6 Brief summary of the chapters covered in the thesis

The format and order of the coming chapters in this manuscript evolve in concordance with the expected structure for a scientific project description.

Chapter 1: Introduction. The topic and the context of the study, which is an omic integration approach to study the relevance of histone modifications on the regulation of the gene expression in the YMC, are presented.

Chapter 2: Materials and methods. The data (RNA-Seq and ChIP-Seq), the experimental design, and all the methodology applied in order to achieve the main objective of the project are described. The description of the statistical methods for the omics integration (MORE and N-PLS) can be found.

Chapter 3: Results. The results obtained with the help of the methodology explained in Chapter 2 are given. We can find here: the clustering and differential expression analysis of the genes from RNA-Seq data, the number of genes significantly regulated by each histone modification and transcription factor, the selection of the most relevant transcription factors involved in the regulation of the YMC, and a regulatory network created to depict graphically the relevance of the different regulators studied, among other results.

A detailed interpretation and discussion of the results obtained in Chapter 3 is given in **Chapter 4: Discussion.** Evidences in the literature that support and strengthen the results are also given.

Last, **Chapter 5: Conclusion,** which summarizes the most relevant aspects of the discussion, the major difficulties faced along the project and future tasks that should be done.

2. Materials and methods

2.1 Omics data analysed

The omics data types used in this project were gene expression, measured with RNA-Seq technology, and histone modifications, measured by ChIP-Seq. All data sets were published in (Kuang, et al., 2014) and retrieved from GEO repository (Edgar, et al., 2002), data's accession number: GSE52339. The only exception was the data of the histone modification H3K18ac which was provided by Dr. Jane Mellor lab, more details in Section 2.1.2.

Metabolic-cycle experiments were performed as previously described in (Tu, et al., 2005) except for the timing of sampling for RNA-Seq and ChIP-Seq; samples were intentionally taken unevenly to more deeply sample the very rapidly changing OX phase and were taken less densely outside the OX phase (Kuang, et al., 2014). Since 16 is the total amount of sampling time points (obtained along one cycle of the YMC) the observations will be named regarding its sampling obtaining order, hence, the first one is 1 or t1 and the last one is 16 or t16.

2.1.1 RNA-Seq

For RNA-Seq, the locations of the 16 sampling time points along the YMC are shown in the Figure 1 from (Kuang, et al., 2014).

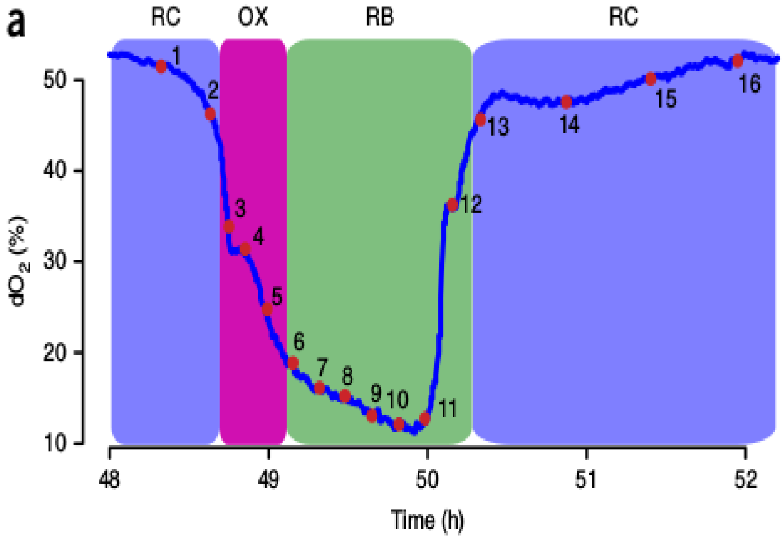


Figure 1. Fluctuation of the oxygen levels (dO_2) along the YMC. The 16 time points of the RNA-Seq data obtaining are labelled in one cycle of the YMC. The three metabolic phases of the YMC are colour coded: magenta for OX, green for RB and blue for RC. Figure from **(Kuang, et al., 2014)**.

RNA-Seq data was sequenced with Illumina HiSeq 2000, the reads were single end and their length was 50 base pairs (bp).

2.1.2 ChIP-Seq

For ChIP-Seq, the locations of the 16 sampling time points along the yeast metabolic cycle are shown in the Figure 2 from (Kuang, et al., 2014).

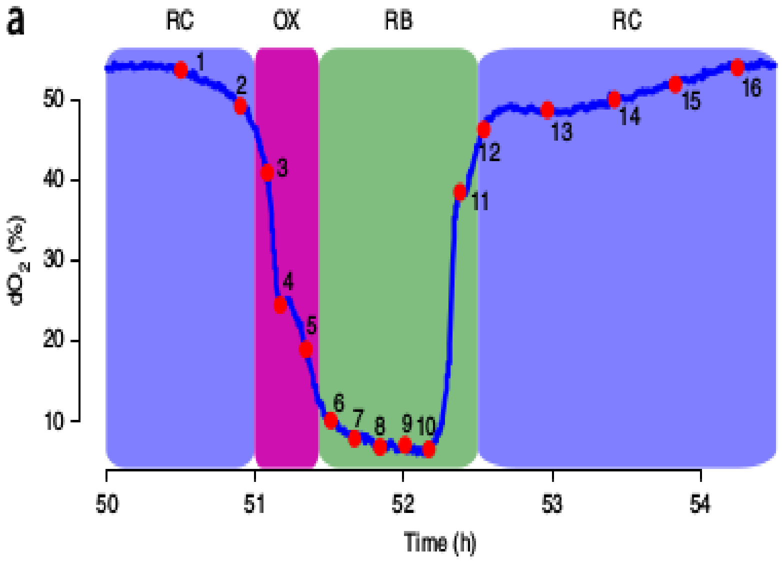


Figure 2. Fluctuation of the oxygen levels (dO$_2$) along the YMC. The 16 time points of the ChIP-Seq data obtaining are labelled in one cycle of the YMC. The three metabolic phases of the YMC are colour coded: magenta for OX, green for RB and blue for RC. Figure from **(Kuang, et al., 2014)**.

ChIP-Seq data was pre-processed from fastq files containing raw sequencing reads.

The Chip-Seq data used in this study corresponded to 8 histone modifications (H4K16ac, H3K36me3, H3K4me3, H4K5ac, H3K9ac, H3K56ac, H3K14ac and H3K18ac), and H3 was used as control. The data had been generated with the usage of three different sequencing technologies, and the average read lengths were different attending to the sequencing technology used (Table 7). The omics data were generated in two different batches, H3K9ac was measured twice, each one in a different batch, in order to use it afterwards to detect and correct the batch effect (this will be discussed in more detail in Section 2.4.3). In the section 1.1 it was indicated that the data used in this project was coming from (Kuang, et al., 2014), however, it must be specified that H3K18ac was not considered and used in their project. Despite that they obtained the aliquots for it in each of the corresponding time points, as for the rest of histone modifications, they did not sequence it. A collaborator of the Genomics of Gene Expression group, Dr. Jane Mellor, had obtained and sequenced the aliquots of H3K18ac. Dr. Jane Mellor gave us the data, so, new information was also added in this study with respect to (Kuang, et al., 2014).

Table 7. Technical details regarding the sequencing technology, average read length (in base pairs) and type of read for each of the histone modifications studied.

Histone modification	Sequencing Technology	Average Read Length (bp)	Type of read
H3K9ac	Illumina HiSeq 2000 ChIP	50-51	Single end
H3K36me3	Illumina HiSeq 2000 ChIP	50-51	Single end
H3K4me3	Illumina HiSeq 2000 ChIP	50-51	Single end
H4K5ac	Illumina HiSeq 2000 ChIP	50-51	Single end
H4K16ac	Illumina HiSeq 2000 ChIP	50-51	Single end
H3	Illumina Genome Analyzer ChIP	36	Single end
H3K56ac	Illumina Genome Analyzer ChIP	36	Single end
H3K9ac	Illumina Genome Analyzer ChIP	36	Single end
H3K14ac	AB SOLiD System	35	Single end
H3K18ac	Illumina HiSeq 2000 ChIP	50-51	Single end

2.2 ChIP-Seq data processing

2.2.1 Reads quality study

A study of the sequencing reads quality was performed with FastQC software (Andrews S, 2010). FastQC is designed to perform different quality control checks on raw sequence data coming from high throughput sequencing pipelines through a modular set of analyses.

In total, 160 ChIP files were analysed (10 histone marks, Table 7, multiplied per 16 time points).

The reads quality was codified based on phred +33 quality scores.

2.2.2 Reads trimming and filtering

Sequencing reads may require a quality filtering to discard those with very low quality before the alignment step and, sometimes, they have also to be trimmed because they may contain adaptor sequences. In order to perform this filtering and trimming, the software Trimmomatic (Bolger, et al., 2014), version 0.32 was used.

For the read files that were not presenting quality problems the parameters used were:
-LEADING:30 TRAILING:30 SLIDINGWINDOW:5:30 MINLEN:28

For the read files that were presenting sequence quality problems, a slightly less restrictive filtering and trimming was applied to avoid discarding the majority of the reads. The parameters used were:
-LEADING:25 TRAILING:25 SLIDINGWINDOW:5:25 MINLEN:25

2.2.3 Mapping, sorting and removing duplicated reads

The reference genome was obtained from Ensembl (Zerbino, et al., 2018) (*Saccharomyces cerevisiae* genome from release 91). To perform the alignment of the filtered and trimmed reads to the reference genome, bowtie1 software was chosen (Langmead, et al., 2009).

Bowtie1 was applied specifying the parameter -m1, in order to discard multimapped reads.

Attending to the nature of the ChIP-Seq data, bowtie1 had to be applied in a different way for two different groups: one group formed by all the samples coming from Illumina technology and another group formed by the samples coming from the Solid technology.

Since the reads from AB solid are codified with the so called “color space” codification, the reference genome should be converted also to color space coding before performing the alignment of those reads, this process was performed also by bowtie thanks to the bowtie-build function, which indexes the genomes, with the parameter -C.

After the mapping, the resulting bam files were sorted and the duplicated reads were removed with the samtools software (Li, et al., 2009).

2.2.4 Averaged number of reads per region

In order to obtain ChIP-seq quantification values, it was first calculated the coverage per nucleotide for all the genome with the program genomecov, from bedtools (Quinlan & Hall, 2010), specifying the parameter –d (this opion returns the coverage per nucleotide).

Next, we defined 20 consecutive non-overlapping regions from the transcription start site (TSS) of each gene. Half of these 20 regions had a length of 100 base pairs (bp) and were upstream the TSS. To define the remaining 10, we divided the gene (from the TSS to the transcription termination site, TTS) into 10 regions with the same length (which also resulted in around 100 bp in median). For each genomic region, we computed the average of the coverage across the region. This was done with a set of python in-house scripts, and with the information of the yeast genome annotation file (gtf). The gtf was obtained from Ensembl (Zerbino, et al., 2018), *Saccharomyces cerevisiae* gtf from the release 91.

Finally, it was analized the average distribution of reads per region for all the genes in 3 of the 160 ChIP-Seq data sets in order to select the regions showing a higher enrichment with respect to the control (H3), the need of using the control is detailed in Section 2.2.5. The selected regions were then used in all the subsequent analyses.

Since two regions were selected, two matrices were created for each histone modification, each of them containing the information of one of the two regions. Each matrix row contained the information of a gene, and each column the information of a different time point.

2.2.5 Normalization

It is certainly relevant taking into account the control ChIP-Seq samples, because the regions presenting high read counts do not necessarily contain enrichment sites. Many studies have described that the distribution of the reads is far from uniform, and is affected by many factors, including GC content, mappability, chromatin structure and copy number variation, among others (Liang & Keleş, 2012).

In order to normalize the histone modifications data with respect to the control, H3, different steps were applied. Firstly, the obtained means for the different studied regions were divided by the total number of reads of the correspondent ChIP-seq sample (sequencing depth correction). Secondly, a millionth unit was added to all the resulting means. Lastly, a log ratio transformation of the means with respect to its corresponding control time point was applied. For instance, the data coming from H3K9ac time point 3 was logically normalized with respect to the H3 time point 3.

2.3 RNA-Seq processing

The RNA-Seq data, obtained from GEO repository, had already been pre-processed by (Kuang, et al., 2014), so this task was not executed in this project. Briefly, the reads were aligned with bowtie1 to the reference genome. Afterwards the data was normalized by the total number of aligned reads per sample, a unit added to the normalized number of reads and a logarithmic transformation applied. Lastly, the different values for each gene were centered.

2.4 Statistical methods

2.4.1 Exploratory analysis

Principal Component Analysis (PCA)

PCA is a mathematical algorithm which enables the reduction of the number of variables (the dimension) in a study, through the creation of new latent variables, also known as components, which retain most of the variability present in the original data matrix (Ringnér, 2008). Each principal component (PC) is a linear combination of the original variables and explains a percentage of the data's variance, in such a way that the first principal component (PC1) explains a larger variance than the second (PC2) and so on, and with a reduced number of those components is enough to explain most of the original variance. The loadings are the weights of the original variables in each of the PCs. Likewise, the weights of the observations over the components are also calculated, and they are named as scores. The projections of the observations and the variables in the new components can be represented graphically, and this kind of graphs contribute to a better comprehension of their relationships and serve as a quality control to check that the observations corresponding to the same experimental groups are clustered together, or if the way in which those groups are separated is consistent with what is expected.

When applying multivariate methods, such as PCA, to obtain a non-biased data representation, it is sometimes necessary centering and or scaling the variables in order for them to present equal average and or similar variance. Generally, it is quite common to center the data, while scaling is usually applied in the case that the variables are measured in different units.

PCA was applied to explore the different omic data sets, and pca function from MixOmics R package (Rohart, et al., 2017) was used. The variables were always centered but not scaled in the PCAs shown along this project.

2.4.2 Differential expression analysis and clustering

2.4.2.1 MaSigPro

MaSigPro is a regression based approach designed to analyse time-series microarray or RNA-seq data (Conesa, et al., 2006; Nueda, et al., 2014). The goal is finding genes whose expression changes with time, and also those with significantly different expression profiles between experimental groups, so a regression model is fitted per gene.

In this work, maSigPro R package was applied to RNA-seq data to obtain the differentially expressed genes (DEGs) across time and their clustering. For the consideration of the DEGs: the significance level was set to 0.05, the p-values were adjusted with the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), and a minimum R squared value of 0.6 was required (more details regarding the biological criteria followed to fix the R squared threshold are given along the Section 3.4).

When applying maSigPro, the user must choose the degree of the polynomial relating time and gene expression. Different degrees were tested and, in order to establish the optimal one, two criteria were applied on the resulting DEGs selected by maSigPro:

- The amount of variance explained by the two components of a PCA on the DEGs data that better ordered the time points, and the graphical quality of this ordering.

-The adjusted R squared values of the different regression models for each gene, which were not returned by maSigPro but computed using Equation 1.

Using the R^2 value to compare regression models with a different number of predictors is not a good procedure, because adding a variable to a model always increases the R^2. Thus, the regression models fitted with higher polynomial degrees will always present larger R^2 values.

To correct the inability to compare models with a different number of predictors using the R^2 value, the adjusted R^2 value was defined. The adjusted R^2 (*R^2adjusted*) will only increase if the added predictor is useful for the explanation of the residual variance present in the model, and it is calculated as indicated in Equation 1:

$$R^2 adjusted = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \qquad \text{(Equation 1)}$$

where n is the number of observations, and p the number of regression parameters (the β coefficients).

2.4.2.2 Genes clustering evaluation

As previously indicated, maSigPro was used to cluster the DEGs, k-means method was chosen.

In order to assess the quality of these clusters, a silhouette distance study was performed (Rousseeuw, 1987). This methodology aims to show the tightness or separation of the different clusters, through a distance measure called silhouette which takes into account the average distance of a particular variable with the rest of variables of the cluster where it has been assigned, and the average distance of this variable with all the other variables from a neighbor and different cluster.

As it is described in (Rousseeuw, 1987) the silhouettes indicate which variables lie well within their clusters and which ones are merely somewhere in between them. The whole clustering is displayed through the combination of all the silhouettes into a single plot, permitting an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity. A silhouette value for a variable close to one indicates that it has been well clustered, while close to zero or negative values indicate the opposite.

2.4.3 Batch effect correction

One of the most complex issues surrounding high-scale omics data analysis is the technical noise (Nikolsky, et al., 2009). It is advisable to remove this technical noise or unknown variance which is associated to the measurement tools or experimental design, because it masks the biological effects which really want to be studied. In order to remove this noise different methods have been developed, being ARSyN (Nueda, et al., 2012) one of them. Through the combination of the analysis of variance (ANOVA), and the subsequent application of multivariate analysis (Simultaneous Component Analysis SCA) over the effects decomposition of the ANOVA, the ARSyN method is able to detect the noise present in the data and correcting it.

RNA-Seq and ChIP-Seq of histone modifications were generated in two different batches, named here as A and B (Table 8). In order to be able to determine the magnitude of the batch effect, the histone modification H3k9ac was independently measured in both batches (Table 8). Thus, regarding H3K9ac different approaches were tested to correct the batch effect. One of them consisted on the usage of the ARSyNseq function from the R package NOISeq (Tarazona, et al., 2015), the other one, simpler, was based on centering independently each H3K9ac data matrix and analyzing afterwards the implications on the batch effect. The reason for the centering approach appears because the RNA-Seq had already been pre-processed, including a centering step which could not be undone, as a consequence it was also planned to center all the histone modifications data before starting with the omics integration.

Table 8. Batch specification for each omic data set.

Omic Data Set	Batch
H3K4me3	A
H3	B
H4K5ac	B
H3K56ac	B
H3K14ac	A
H3K9ac	A, B
H3K36me3	A
H4K16ac	A
H3K18ac	B
RNA-Seq	A

2.4.4 Omics Integration

2.4.4.1 Time point alignment

Analysing Figure 1 and Figure 2 , it can be seen that the exact number of time points per metabolic phase and their relative position regarding the levels of oxygen consumption are not the same for RNA-Seq and ChIP-Seq data. In the RB phase for the RNA-Seq data there are 7 sampling points, and 6 for the ChIP-Seq data. In the RC phase, when it comes after the RB phase (right side of the image), there are 5 time points for the ChIP-Seq data and 4 for the RNA-Seq data.

Since the omics integrative analysis requires having matchable observations from the different omics, this disagreement between time points must be fixed. Taking the oxygen consumption levels as an indicator of the metabolic state of the cell, and using them as a reference, it was decided to average the data from time point 10 and 11 of RNA-Seq, and to average the data from the time points 13 and 14 of the ChIP-Seq. As a consequence, after these adjustments 15 time points were left, instead of the initial 16.

2.4.4.2 N-PLS

In Principal Component Analysis (PCA), as pointed in Section 2.4.1, the aim is the reduction of the number of variables which describe the data variance through the creation of latent variables, to facilitate the comprehension and analysis of the changes in the observations. In general, in the genomics field, only one type of omic data is represented in a PCA, for instance RNA-Seq data. When dealing with different kinds of omics data simultaneously (in our case, RNA-Seq and ChIP-Seq data), the common objective in the multivariate modelling is going to be also the reduction of the dimensionality in each kind of data, while maximizing the relationship between the two types of omic, the covariance. In order to accomplish that, a widely used multivariate methodology is the Partial Least Squares Regression (PLS) (Wold, et al., 2001), where one of the data matrices is considered the response variable Y (e.g. RNA-seq) and the other is the prediction matrix X (e.g. ChIP-seq).

However, in our case, we do not have a single ChIP-seq prediction matrix but many of them, since we have performed ChIP-seq experiments for several histone modifications and, in addition, we have quantified them for two different genomic regions. Therefore, the X prediction matrix with genes in rows (first mode) and observations (time points) in columns (second mode) becomes a three-dimensional structure where the third mode (slices in Figure 3) represents the different histone modifications per genomic region. To analyse such three-way structured data, several possibilities exist, one of them consisting of unfolding the cube and convert it into a matrix by concatenating the different slices. However, specific methods have been developed to deal with the three-way problem, as for instance the N-PLS model (Bro, 1996), that was used in this project to integrate RNA-Seq with all the ChIP-Seq data sets.

To facilitate the comprehension of what N-PLS methodology implies, first of all, let us review the PLS model (Wold, et al., 2001), since N-PLS is the natural extension of PLS to N-way structures (Conesa, et al., 2010). PLS can be considered as a prediction model, and it requires two data matrices as explained above: the matrix of the predictors and the matrix of the responses. The aim of the model is being able to predict, in the best possible way, the response matrix from the predictor matrix, based on the common behaviour or structure of the variables of both matrices. In order to do that, when obtaining the latent variables or components, the model aims to maximize the explained variance for each of the components attending to the location in the space of the observations from both matrices, maximizing thus the covariance.

The N-PLS method, though, is considered as a multiway data analysis strategy. Multiway data analysis includes a number of methods developed to analyse large data sets through the representation of the data as multidimensional arrays (more than two dimensions**)** (Coppi, et al., 1989). Multiway techniques could be considered as extensions of traditional bilinear dimension reduction techniques to multidimensional data structures (Smilde, et al., 2004). For instance, the multiway Tucker3 (Smilde, et al., 2004) method can be considered as the extension of PCA to multidimensional structures. When dealing with multiway matrices each dimension is referred to as modes. Thus, in a 3 dimensional matrix, such as the ChIP-seq data matrix of this project, there are 3 modes.

To interpret the results rendered by the N-PLS or Tucker3 techniques, the G matrix must be analysed, also called the core matrix. The analysis of the core matrix provides the information of which combination of components, one per mode, are the most relevant because they describe most of the variance present in the N-dimensional data.

In short, the N-PLS method constitutes a dimensional reduction multivariate technique which takes profit of the multiway methodology to decompose the variance of matrices with more than 2 dimensions, together with the usage and extension of the PLS modelling, which maximizes the covariance of the created latent variables when decomposing the variance of two matrices.

In this project, RNA-Seq data was taken as the 2-way response matrix in the N-PLS model, while the ChIP-Seq data was considered the 3-way predictor matrix, since the objective was modelling gene expression changes based on the chromatin status data.

Regarding the modes, and the structure of the matrices, as commented before, the genes were located in the first mode, the time points in the second mode, and the third mode was for the different histone modifications per genomic region (Figure 3).

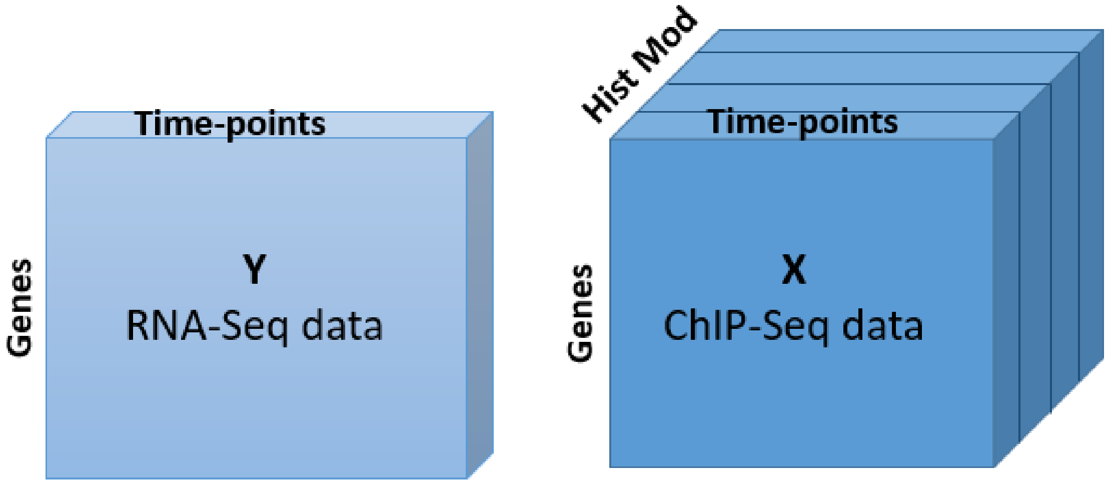


Figure 3. Data structure for the N-PLS model.

To perform the N-PLS an in-house R package was used. The matrices were centered along the mode 2, per time point.

2.4.4.3 MORE strategy

MORE (Multi-Omics Regulation) R methodology (Genomics of Gene Expression group, s.f.) is being developed by the Genomics of Gene Expression group. MORE is mainly based on regression models that aim to explain gene expression (or similar variables) as a function of regulatory elements.

The objective of MORE is the identification of relevant regulators from a large subset of putative regulators, finding those which are significantly modulating the levels of a response variable, which could be: gene expression, protein or transcript levels, among others. MORE is based on the usage of generalized linear models (GLM), developed by (Nelder & Wedderburn, 1972). GLMs extend the linear models, which assume a gaussian distribution for the response variable, to allow for a broader set of statistical distributions belonging to the exponential family: Normal, Binomial, Poisson and Gamma, among others. In GLMs, the regression coefficients are no longer estimated by means of the least squares approach as in linear models, but by the maximum likelihood method. Hence, MORE is able to model the behavior of many different data types, due to the flexibility given by the GLM: normally distributed response variables (e.g. microarray expression data), as in classical linear regression or count data (e.g. RNA-Seq expression data) that can be modeled with the Poisson or the Negative Binomial distribution, the latter being preferred to account for overdispersion.

As pointed before, MORE is designed to evaluate the significance of the effect of a subset of putative regulators on the levels of a response variable. In this project, the response variable was gene expression. For each of the studied genes (the DEGs in our case), a model is created. The predictor variables are the potential regulators, in our case, either histone modifications or transcription factors. Both, the response and the regulators, must have been measured on the same subjects (time points in our case). Besides the potential regulators, the experimental descriptors (treatment, types of tumor, etc.) can be optionally added as covariates. In this project, there were no experimental covariates.

But, together with the GLMs, many more functionalities are available in MORE (Figure 4) in order to facilitate the integration task to the users. Some of them are described next.

Each gene has a different number of potential regulators, so MORE will automatically generate the initial model equation. For instance, a TF can regulate, generally, the expression of a set of target genes, but it can not act over all the genes being expressed. MORE just requires that the user specifies the association between the regulators and their targets, so an association matrix must be provided.

It is quite common to have more predictors in the initial model equation than observations in the data. To tackle this problem, MORE offers several options that can be applied independently for each omic data type, as for example a low variability filtering, a multicollinearity filter, and two variable selection strategies, which are penalized and stepwise regression.

As described in (Faraway, 2005) multicollinearity arises when some predictors are close to be linear combinations of others. It leads to imprecise estimates of the beta coefficients of the regression model, and the signs of the coefficients can be the opposite of what instinct about the effect of the predictor might suggest. The standard errors are inflated and as a consequence the tests may fail to reveal significant factors. As a result, the fitted model is very sensitive to measurement errors, where small changes in the response variable can originate large changes in the estimated beta values. Thus, multicollinearity is a serious problem that must be faced. The multicollinearity filter strategy applied by MORE consists in computing the correlation between the different predictors, and aggregating highly correlated predictors by either averaging their values or randomly choosing a representative regulator among them to be included in the model. Therefore, this implies that the significance of all the predictors, in a group of highly correlated variables, will be assessed by just one beta coefficient.

Regarding the variable selection approaches, the penalized regression implemented in MORE is the Elastic Net shrinkage approach (Zou & Hastie, 2005), which combines Ridge (Hoerl & Kennard, 1970) and Lasso (Tibshirani, 1996) strategies. For the stepwise variable selection (Draper & Smith, 1998), several procedures were adapted in MORE: forward, backward, two ways forward and two ways backward.

Additional functionalities in MORE are a summary of the resuls at global level and at gene-specific level, and plots displaying the relationship between a given gene and its significant regulators according to the GLM results.

As commented before, MORE was applied in this project to evaluate the relevance of different histone modifications and transcription factors in the regulation of the differentially expressed genes.

In order to determine which genes were transcription factors, and which their target genes were, Yeastract (Teixeira, et al., 2018) data base was used.

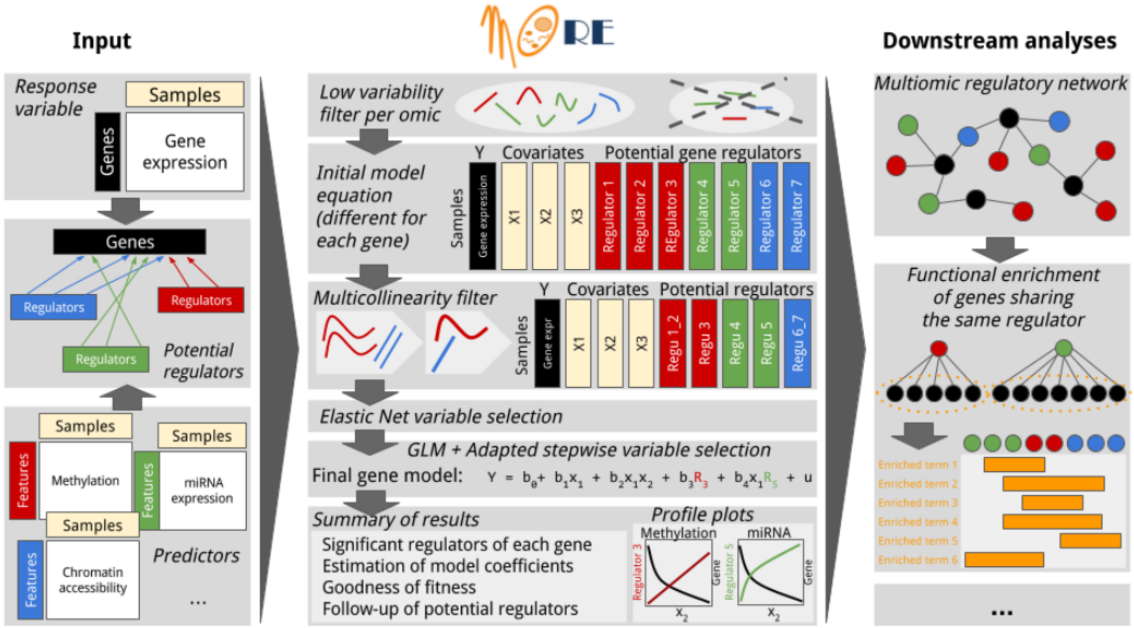


Figure 4. MORE method overview. The diagram represents the different steps of the analysis and some ideas for the downstream analysis that can be done from MORE results.

2.4.5 Functional enrichment analysis

Functional enrichment (FE) studies are procedures inspired in the systems biology criteria. The aim of these approaches is the direct examination of functionally related groups of variables, like genes or proteins, instead of studying them individually (Dopazo, 2006). In addition, it must be taken into account that the extraction of the biological information is not an easy task, to face this issue a specific vocabulary to annotate the function of the different biologic variables (proteins, genes...) has been developed in a systematic way (Zhou & Su, 2007) which allows to access them in a fast and standarized way, the Gene Ontology (GO) notation system is a perfect example. The GO notation establishes an organized system, surrounding a hierarchical structure, which defines a series of descriptive terms for the different biological entities in 3 different aspects: biological processes, molecular function and cellular components (Zhou & Su, 2007). Thus, for instance, it can be retrieved the GO terms associated to a given protein identifier, and then know in which biological processes it is involved, its biological functions or whether it is a part of a cellular component. In this way, thanks to these standarized GO terms it is possible to find, for example, two different proteins with the same functions. Taking profit of this, functional enrichment analysis can be applied, in order to detect whether there are significant differences in terms of the relative abundance of a specific function between 2 groups of variables.

Thus, functional enrichment analysis is performed on different established groups, and it is equivalent to a test for independence of two variables: belonging or not to the test set of variables (e.g. differentially expressed genes) and being or not annotated to a given functional term. Specifically, Fisher’s exact Test was applied here to find statistically significant differences in terms of function abundance between groups. The significance level for these analyses was set to 0.05. The p-values were adjusted with the (Benjamini & Hochberg, 1995) procedure. To perform the functional enrichment analysis an R in-house script was used.

In this work, Gene Ontology (GO) (Gene Ontology Consortium, et al., 2004) and KEGG (Kanehisa & Goto, 2000) data bases were used to retrieve the functionality associated to the different studied genes.

2.4.6 Selection of the most relevant transcription factors

Once MORE was applied to find the transcription factors significantly regulating gene expression, the analysis described below was carried on to select the most relevant transcription factors (TFs) governing the YMC.

For each of the studied TFs, the total number of associations with the DE target genes was calculated from Yeastract database. Many of this associations, despite they had been described in the literature, were not truly happening in the context of the YMC, and that is why the MORE turned out to be a useful tool to filter such associations, by maintaining those which were significantly more likely to be happening. Therefore, the total number of TF-target gene associations according to MORE results was also calculated.

A TF X was considered to be relevant when the proportion of significant regulations (according to MORE) among all the regulations in which X was involved significantly increased in comparison with the rest of regulations (see Figure 5). Thus, in order to test the relevance of TF X, a Fisher's Exact test can be applied.

TF X	**Nr. of genes Regulated by the TF X**	**Total nr. of regulations, given all the studied TFs, excluding the ones given by the TF X**
Significant in MORE	a	b
Not significant in MORE	c	d

Figure 5. Representation of the contingency table used for the Fisher's Exact test of a given transcription factor (TF X). a is the number of target genes for TF X described in the literature which are being significantly regulated by this TF attending to MORE. c is the number of target genes for TF X which are not being significantly regulated by this TF. b is the total number of significant regulations excluding those involving TF X. d is the total number of non significant regulations excluding those involving TF X.

Thus, the relevant TFs were those with a significant adjusted p-value for the Fisher's Exact Test (significance level 0.05), and with an odds ratio higher than 1. Benjamini and Hochberg procedure (Benjamini & Hochberg, 1995) was applied for multiple testing correction.

2.4.7 Significant associations analysis

Another step of the analysis consisted on evaluating if any of the most relevant transcription factors were significantly regulating the same group of DEGs that any of the histone modifications. To do that, a Fisher's Exact test was again applied as indicated in Figure 6.

	Nr. Genes Significant for TF X	Nr. Genes non-significant for TF X
Nr. Genes significant for Hist mod Y	a	b
Nr. Genes non-significant for Hist mod Y	c	d

Figure 6. Contingency table used for the Fisher's Exact test applied to evaluate if any histone modification (Hist mod Y) was significantly regulating the same group of genes that any TF (TF X). a is the number of genes significantly regulated by the studied TF and hist mod. b is the number of genes which are significantly regulated by the studied histone modification but are not by the studied TF. c is the number of genes which are significantly regulated by the studied TF but are not by the studied histone modification. d is the number of genes not regulated by neither the studied TF nor histone modification.

The significant associations were those with a FDR adjusted p-value (Benjamini & Hochberg, 1995) below the significance level of 0.05, and with and odds ratio higher than 1.

2.5 Computer support, software and programming languages

The programming languages used along this master thesis project were: R (R Development Core Team, 2011) version 3.4.3, Python (Python Software Foundation, s.f.) version 3, and bash scripting.

R was the most used programming language, through the usage of the graphical interface R studio.

Python language was used mainly along the ChIP-Seq data processing to generate the genome coverage matrices. It was also utilized to parse and analyze the reads quality, trimming and mapping results.

Bash scripting was used to pre-process the ChIP-Seq raw data.

With the exception of the pre-processing of the ChIP-Seq raw data, all the analysis were run in a computer with GNU/Linux Ubuntu 16.04 software, with 4 cores and 4 Gb of RAM. ChIP-Seq raw data sets were analysed and pre-processed in the cluster from Centro de Investigación Príncipe Felipe.

2.6 Pipeline overview

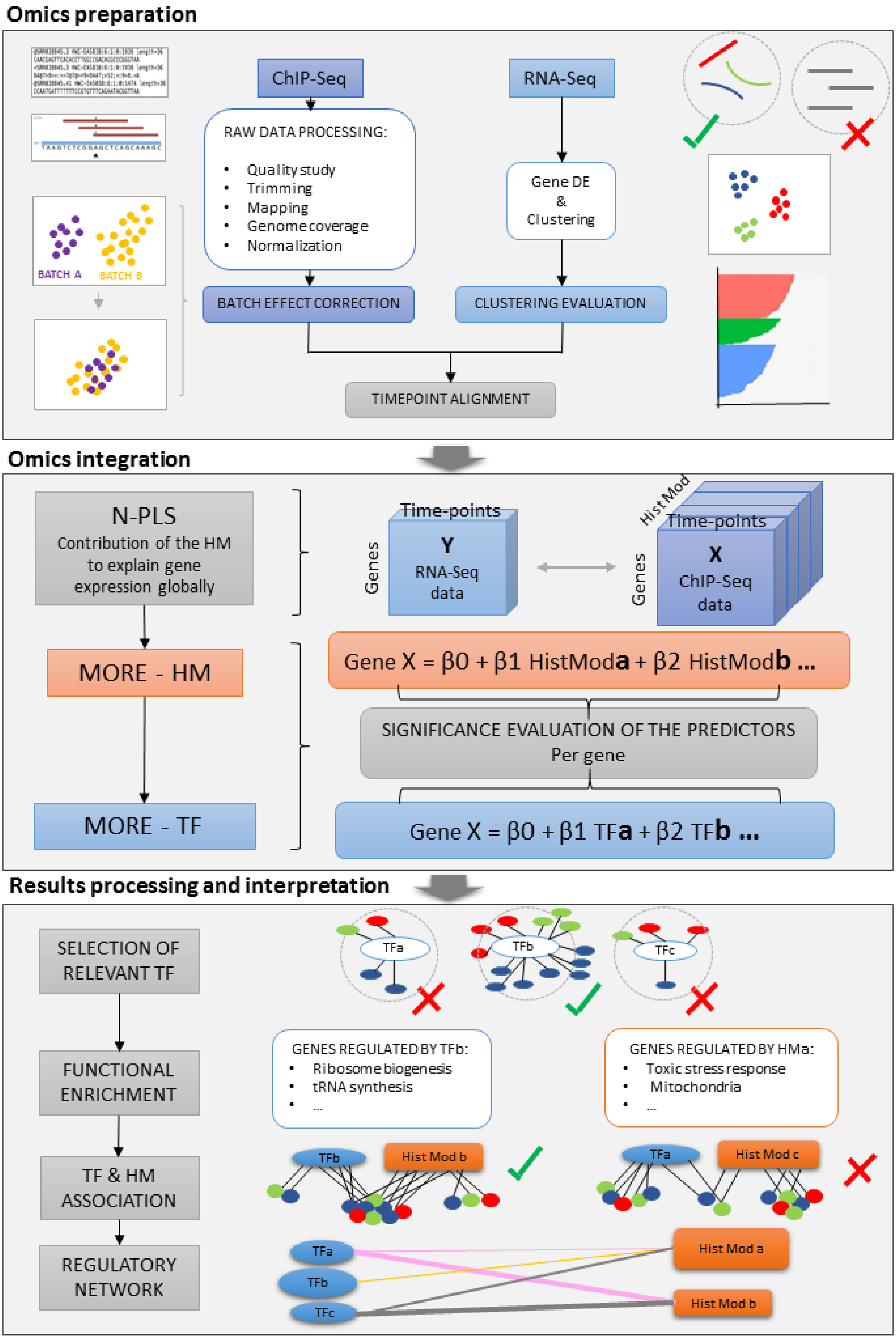


Figure 7. Project pipeline. TF (transcription factor), HM (histone modification).

3. Results

3.1 ChIP-Seq data processing results

3.1.1 Quality study, trimming, mapping and duplicated reads removal

After performing the quality study of the reads with the FastQC software, 28 ChIP-seq samples out of 160 presented a low sequence quality: the 16 time points from H3K14ac histone modification (the only histone modification sequenced with AB SOLid System) and 12 time points from H3K36me3. Figure 8 shows an example of good and bad quality samples.

Based on the read quality differences among the ChIP-seq samples, two different filtering protocols were applied (Section 2.2.2). The least restrictive filtering was applied on H3K14ac and H3K36me3 data, to avoid discarding the majority of the reads. A more restrictive read filtering was applied on the rest of samples.

The FastQC study also indicated a possible presence of adapters in 32 samples that corresponded to all the time points of H4K16ac and H3K4me3. As a consequence, a study of the adapters sequences was performed for the different time points in these histone modifications, and 11 of the most overrepresented sequences were selected (not all of them due to the high similarities between the sequences). These sequences (see Figure 9) were used in the trimming step as putative adapter sequences.

After the reads trimming, the FastQC software was used again to evaluate if the previous problems had been solved. We observed that all the ChIPs presented a good sequence quality and did not present adapters. In the Figure 10, an example of the quality correction is shown for a particular ChIP-Seq sample.

H3K36me3 and H4K16ac were the most affected ChIP-Seq samples by the trimming step. H3K36me3 was one of the samples also presenting sequencing quality problems, and H4K16ac had adapters in the sequences. Along the different time points of these ChIP-Seq experiments, the number of final reads left after the filtering and trimming steps was below one million, being the H4K16ac time point 1 the most extreme case, where just around twenty thousand reads were left. The removal of a notorious amount of reads for the time point 1 in H4K16ac was not surprising, since the FastQC study showed that there was a sequence (adapter) of approximately 42 base pairs (bp) present in more than 75% of reads, and the length of these reads was 50 bp. Probably, something failed in this ChIP-Seq preparation and only the adapter was sequenced. After evaluating the trimming results, it was decided to discard the histone modifications H3K36me3 and H4K16ac from the analysis.

Once the trimming was completed, the remaining reads were mapped to the yeast reference genome with Bowtie 1, and multimapped and duplicated reads were removed. The resulting bam files were used to compute the coverage per nucleotide, i.e. the number of sequencing reads at each nucleotide.

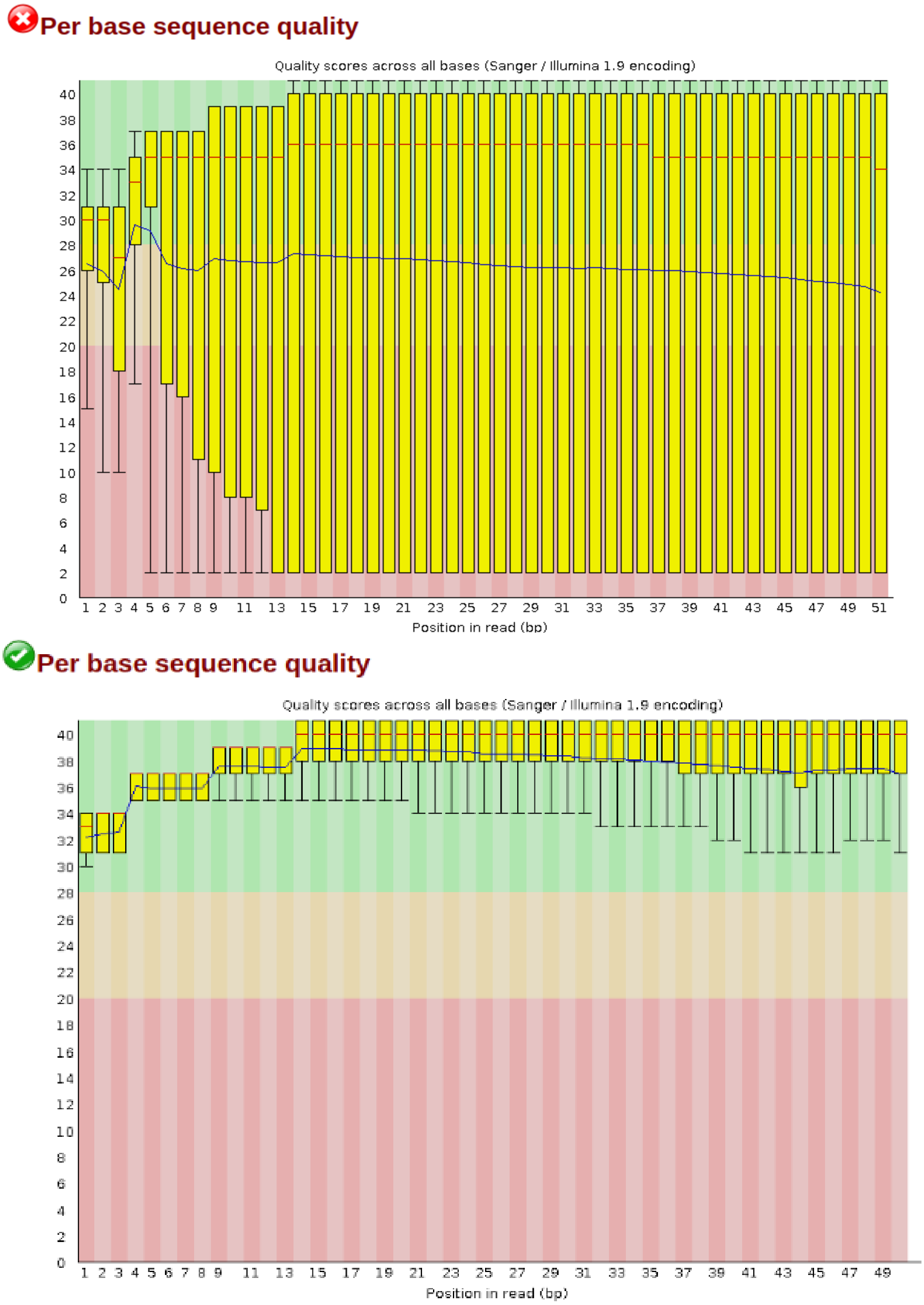
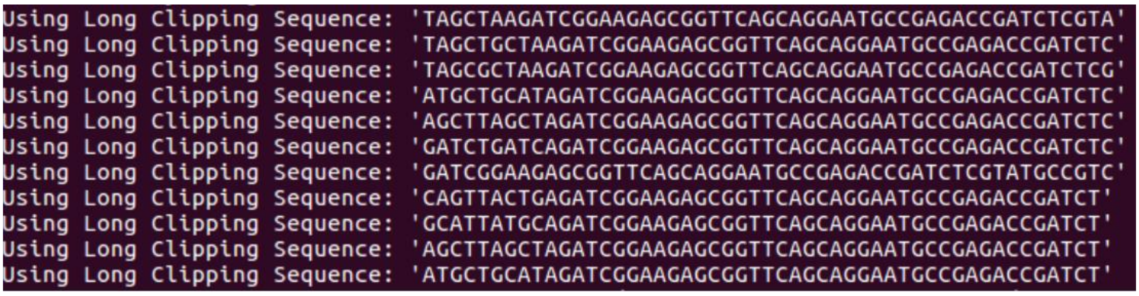


Figure 8. Quality scores across all bases. Example of some of the ChIP-Seq samples presenting a good sequencing quality (bottom) and a bad sequencing quality (top).



```
Using Long Clipping Sequence: 'TAGCTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA'
Using Long Clipping Sequence: 'TAGCTGCTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC'
Using Long Clipping Sequence: 'TAGCGCTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG'
Using Long Clipping Sequence: 'ATGCTGCATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC'
Using Long Clipping Sequence: 'AGCTTAGCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC'
Using Long Clipping Sequence: 'GATCTGATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC'
Using Long Clipping Sequence: 'GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTC'
Using Long Clipping Sequence: 'CAGTTACTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT'
Using Long Clipping Sequence: 'GCATTATGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT'
Using Long Clipping Sequence: 'AGCTTAGCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT'
Using Long Clipping Sequence: 'ATGCTGCATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT'
```

Figure 9. Considered adapter sequences used in the reads trimming.

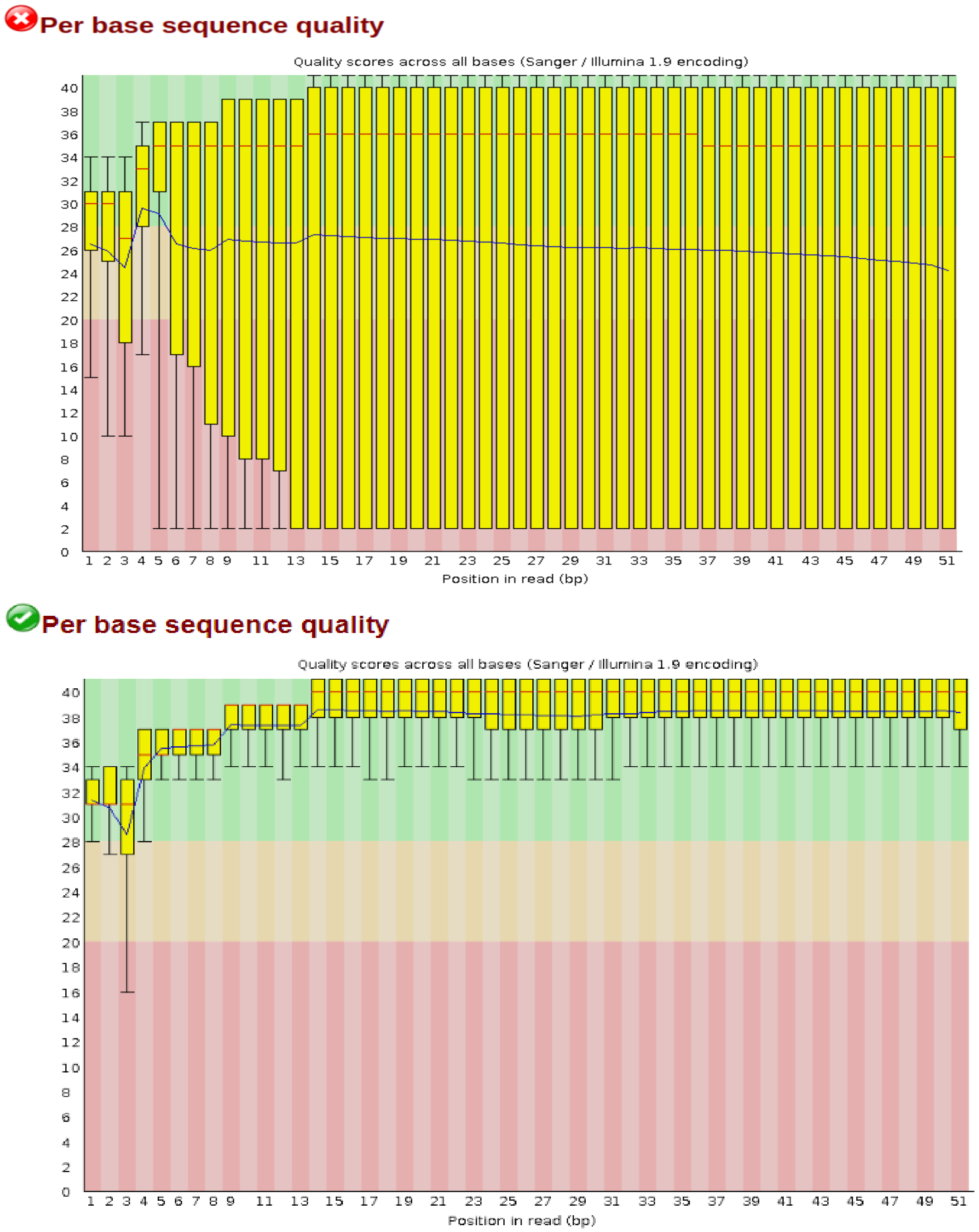


Figure 10. Quality scores across all bases. Example of the reads quality for a particular ChIP-Seq sample before (top) and after (bottom) the trimming.

3.1.2 Determining the genomic regions to be used for the study of histone modifications

Although the coverage per nucleotide along the whole genome can be computed, the most interesting regions to study the chromatin status are those which are close to the genes. However, there is no consensus in the literature on how to define the location or width of such regions, so we performed a study to identify the most informative ones.

Thus, 20 genomic regions were defined for each gene, as described in Section 2.2.4, and the coverage per nucleotide within the region was averaged. This was done for all the genes in 3 of the 160 ChIP-Seq data sets. Figure 11 shows the distribution of the region averages for the 20 regions before and after H3 control normalization (see Section 2.2.5). Since the effect of the histone modifications under study was mainly associated to promoter regions, it was expected to have larger averages close to them. Hence, the distribution of the averages after normalizing with the H3 control was more meaningful and pointed to this expected behavior.

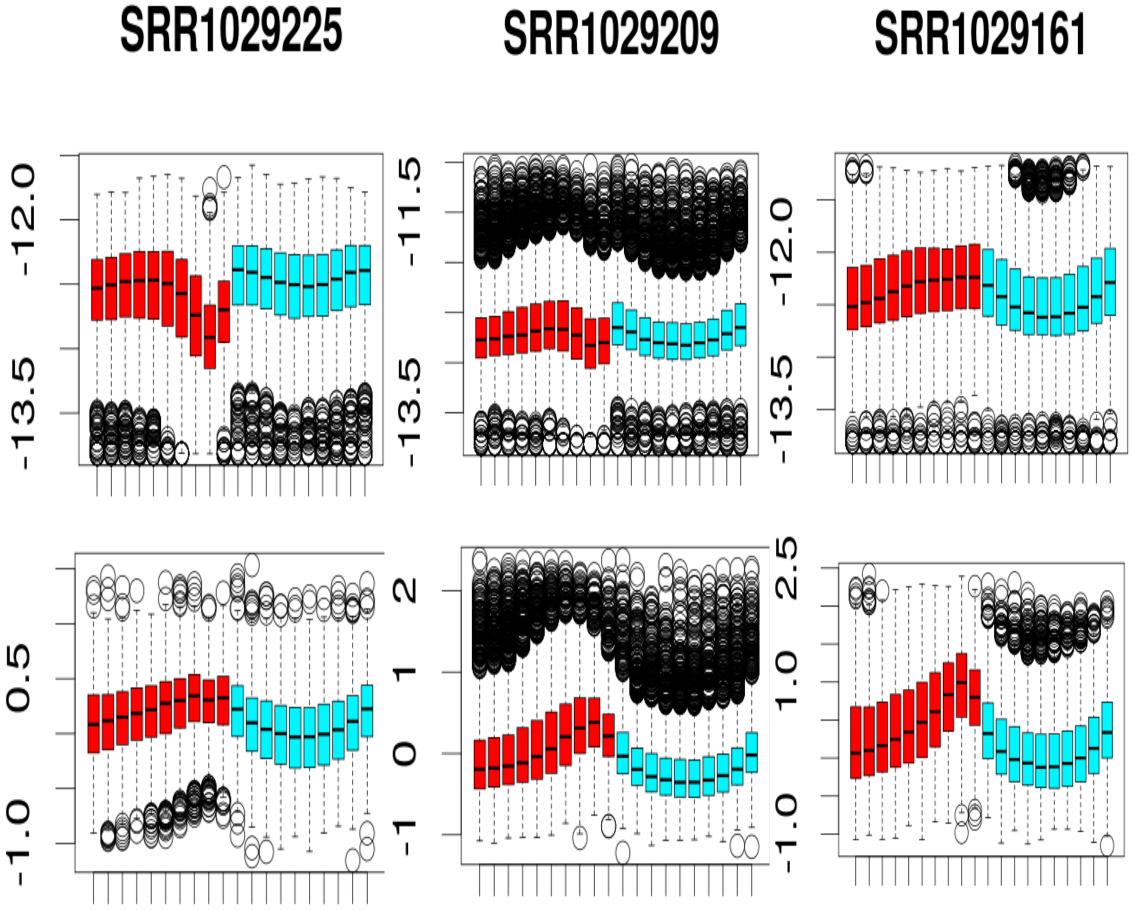


Figure 11. Distribution of the average number of reads along the 20 studied regions for 3 different ChIP-Seq samples. Each column of the figure corresponds to a particular sample. In the first row, averages were not normalized with H3. The averages in the second row were normalized with H3. Red coloured boxplots correspond to regions upstream the TSS, while blue coloured bars correspond to regions in the gene body.

As an additional checking and for the same ChIP-Seq samples, the ten genes presenting the highest variation in their associated histone modification signals were selected, and their behaviour across the 20 regions was studied (Figure 12).

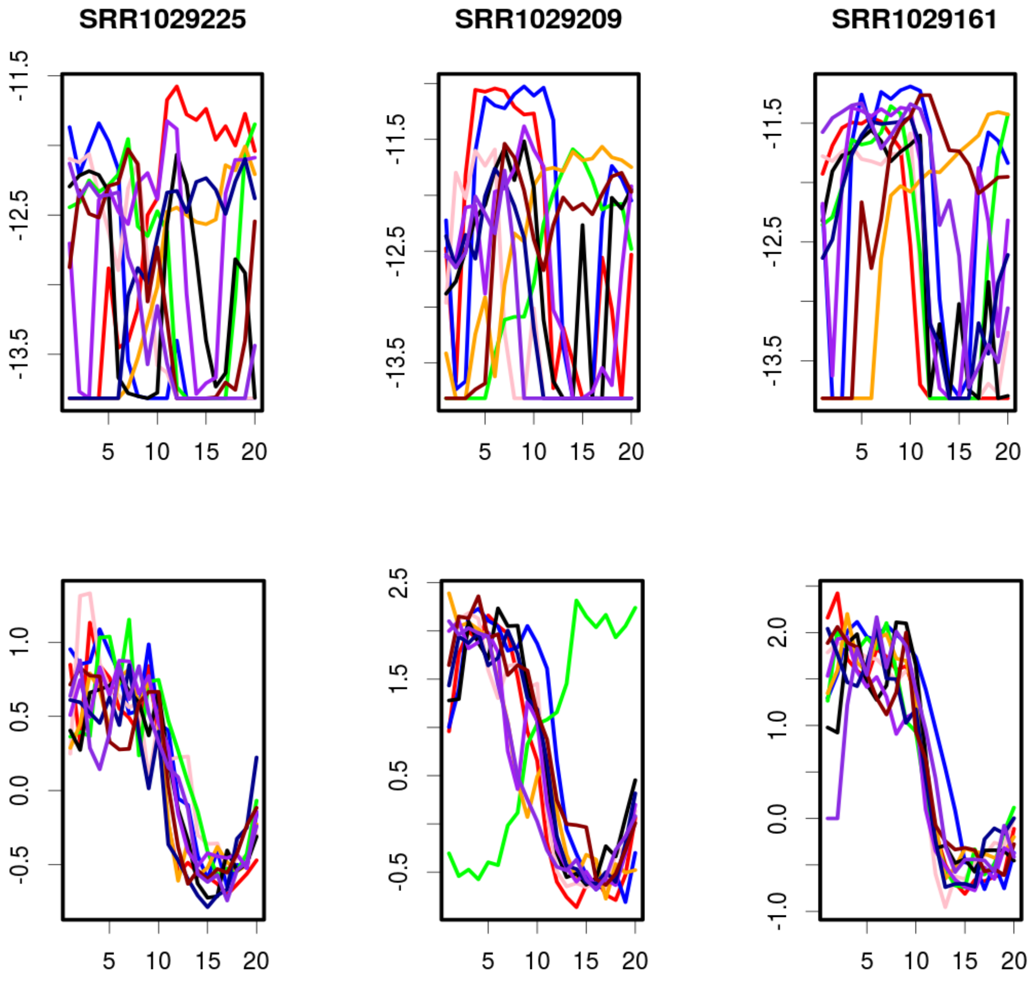


Figure 12. Average distribution of reads for the 20 studied regions in the 10 genes presenting a higher variation for 3 different ChIP-seq samples. Each column of the figure corresponds to a particular sample. The averages of the first row were not normalized with H3. The averages in the second row were normalized with H3.

Before H3 normalization, no relevant patterns could be appreciated in Figure 12, whereas after normalization the expected tendency of having larger means close to the promoter is accomplished. A notorious decrease of the signal is globally observed after the TSS. Again, since the effect of the studied histone modifications is mainly associated to promoter regions the obtained results indicate that the pre-processing was done correctly.

After analysing these results, two regions of 300 bp each were defined, one of them starting 300 bp upstream the TSS and finishing at the TSS, and the other from the TSS to 300 bp downstream the TSS. Therefore, for each histone modification, we will have two different quantification matrices, each one of them containing the values of the average coverage for each of these two regions per gene and time point.

3.2 Batch effect removal

As detailed in Section 2, the different omic data types were generated in two different batches. In order to evaluate the batch effect, the histone modification H3K9ac was measured in both batches. However, it should be noticed that the sequencing platform was different for each batch (Table 7). Although the differences between the sequencing platforms could be really small, the batch effect is confounded with the sequencing platform effect so it cannot be accurately estimated.

Let us recall that the two regions defined to study the histone modification effects were located from 300 bp upstream the TSS until the TSS, and from TSS to 300 bp downstream the TSS, respectively. We analysed the batch effect only for the first region considered.

A PCA analysis was performed on the data from both H3K9ac experiments to assess the relevance of the batch effect. The score plot is represented in Figure 13. It can be appreciated how all the observations are clustered according to the batch they belong to, and not according to the time point as expected, so the batch effect can be perceived. Despite the batch effect was quite obvious from the PCA, we also confirmed it with an analysis of variance (ANOVA) model. The aim of the ANOVA was testing, per gene, the significance of the "batch" factor. The results of the ANOVA, after Benjamini and Hochberg p-value correction, indicated that for 2467 genes out of the 2552 DEGs (97% of the DEGs), the batch effect was significant. Thus, the batch was significantly affecting the histone modification levels.

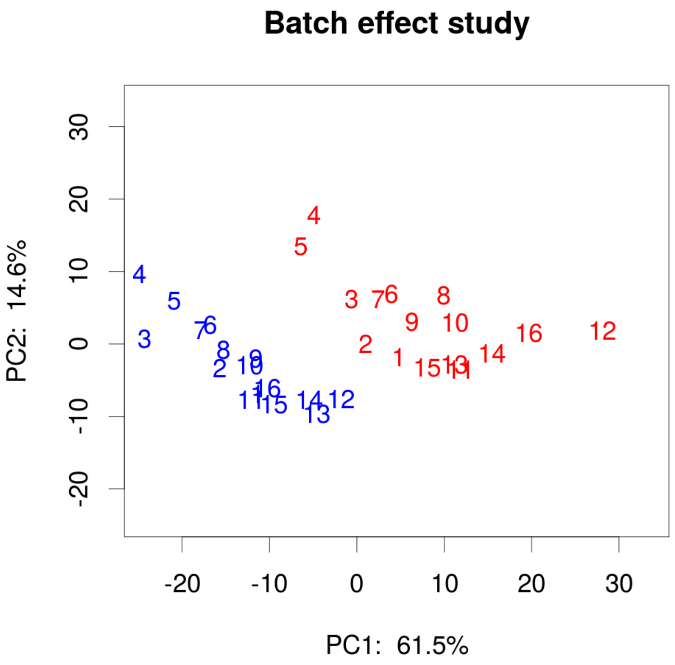


Figure 13. Score plot for the PCA obtained from the data of both H3K9ac experiments. The numbers represent the time points when the samples were taken in the YMC. In red, samples from batch A, and in blue, samples from batch B.

With the objective of correcting the observed batch effect two strategies were followed: The ARSyN method and a centering approach. Since the RNA-Seq data used in this project had already been pre-processed, including a centering step which could not be undone, the same centering approach was followed for the histone modification data for H3K9ac.

Figure 14 (left) shows the PCA score plot for the ARSyN corrected histone modification values. It can be noticed how after the ARSyN correction the observations are no longer separated by the batch and equal time points are clustered together, despite some time points are closer than others. The score plot for the PCA on the centered H3K9ac data is shown also in Figure 14 (right).

Comparing the results of both strategies (Figure 14), and according to the graphical separation of the observations, it is perceivable that centering the matrices is equivalent to applying the ARSyN method, since the score plots are identical. Thus, the effect of the batch in this study was associated with an average increase of the histone modification levels per observation and gene, and can be removed by centering the data per gene.

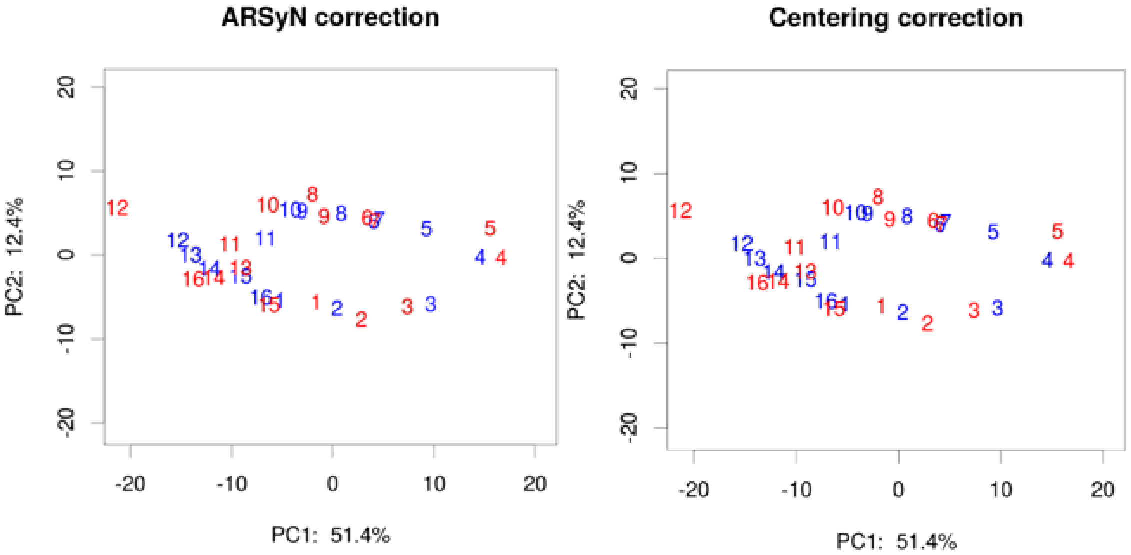


Figure 14. Batch effect removal. Score plot of the PCA obtained after the correction of the data batch effect with ARSyN (left). Score plot of the PCA obtained after centering the data matrices of H3K9ac before merging them into a single matrix (right). In red, samples from batch A, and in blue, samples from batch B.

An ANOVA model on the corrected data confirmed that the batch effect was not significant anymore for any of the 2552 genes. Consequently, this batch effect correction based on centering the data was applied to the data sets of all the histone modifications.

As detailed in Section 2.1.2, H3K18ac was independently sequenced by Jane Mellor, a collaborator who provided us this histone modification data. Despite the sequencing of the samples for this histone modification was performed independently of the rest of the histone modifications considered in the study, the biological preparation for the samples of this histone modification was exactly the same as for samples in batch

B (Table 8), although the sequencing was performed later on. Hence, H3K18ac could present a batch effect, not associated to the sample preparation, but associated to the sequencing time. An exploration with PCA (figure not shown) and the N-PLS results that will be exposed in next sections showed no batch effect, since all the H3K18ac time points were clustered with the corresponding time points in the rest of histone modifications. In case a batch effect had been noticed, it would not have been possible to correct it since there were no replicates of this histone modification in the other batches and we would have had to get rid of this histone modification, but fortunately it could be used in our study. Hence, the batch effect correction applied with H3K18ac was the same applied to the rest of ChIP-Seq samples, the centering approach.

3.3 H3K9ac selection

Since there are two samples of H3K9ac, which will be referred to as H3K9ac1 and H3K9ac2, it must be decided whether averaging them or discarding one of them before the omics integration analysis.

Different aspects were evaluated to take this decision. On the one hand, the amount of remaining reads available for each histone modification per time point was considered, to see which dataset could be more reliable. As it can be appreciated in Figure 15, H3K9ac1 has twice the amount of available reads per time point than H3K9ac2, 8 million per time point (acceptable sequencing depth) in front of around 3 million per time point for H3K9ac2, which is a poor sequencing depth.

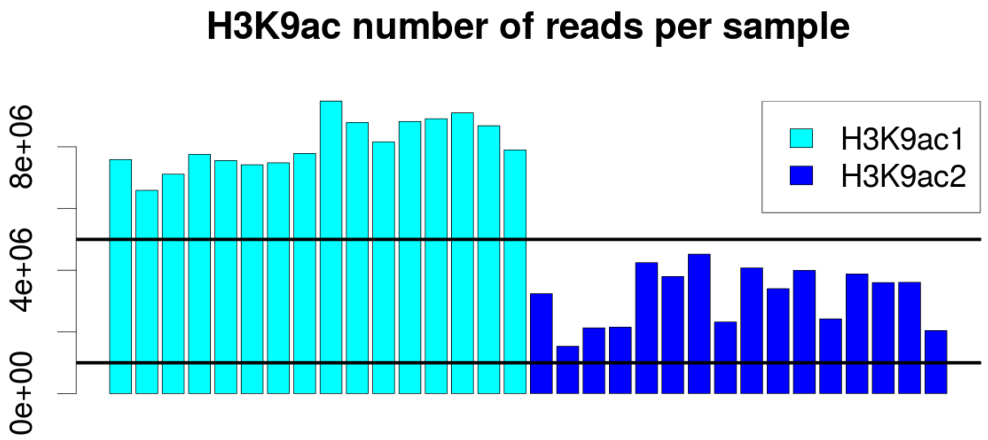


Figure 15. Number of reads left after the ChIP-Seq pre-processing for H3K9ac1 (light blue) and for H3K9ac2 (dark blue). The horizontal black lines represent the height corresponding to one million of reads (bottom) and five million of reads (top).

On the other hand, an independent PCA for each H3K9ac data set show a better performance for H3K9ac1 (Figure 16), since H3K9ac1 provides a better separation of the time points, in a nicer cyclic behaviour, in comparison with H3K9ac2. Moreover, the explained variance with the two first principal components was also higher for H3K9ac1.

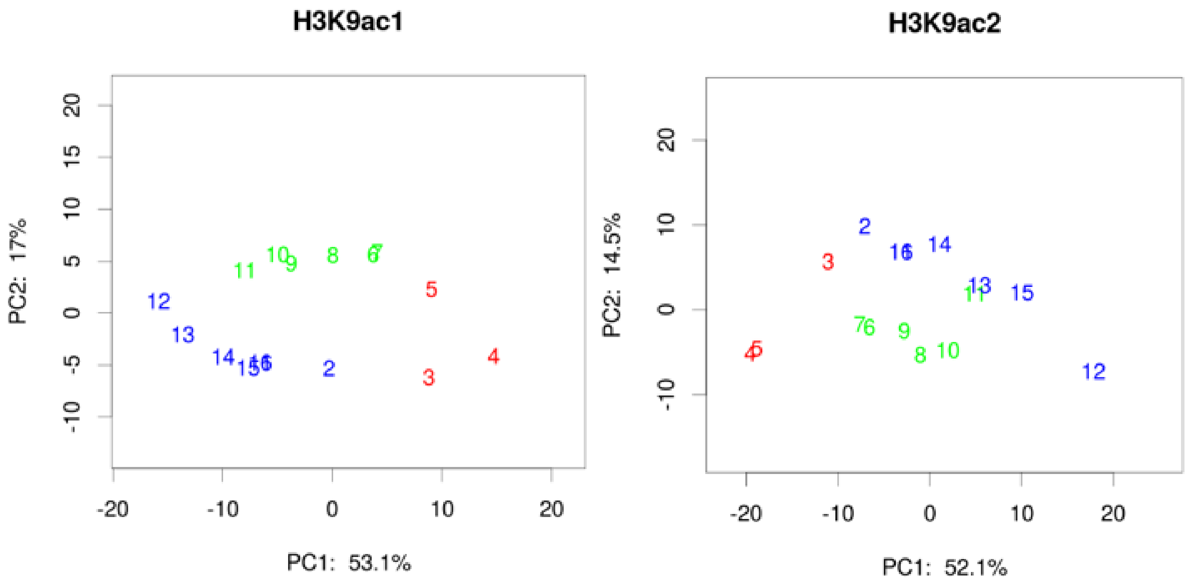


Figure 16. Score plots for the PCAs of each H3K9ac data. H3K9ac1, left, H3K9ac2, right. In green, the time points associated to the RB phase, in blue, the time points associated to the RC phase, and in red, the time points associated to the OX phase.

According to these results, the histone modification named in this section as H3K9ac1 was the one selected for further analyses.

3.4 RNA-Seq data analysis

3.4.1 Initial exploration

The original RNA-Seq data set contained 6035 genes. Out of them, 43 presented all their values equal to 0 and were consequently excluded from the study. A PCA with the resulting 5992 genes was done. Figure 17 shows the scores for the different time points for the PC1 and PC3, the components best separating the time points according to the different metabolic phases of the YMC. It can be appreciated how the time points are distributed in a nice cyclic manner, representing the expected relationship between time points (Figure 1). The time points are clearly clustered attending to the metabolic phase they belong to.

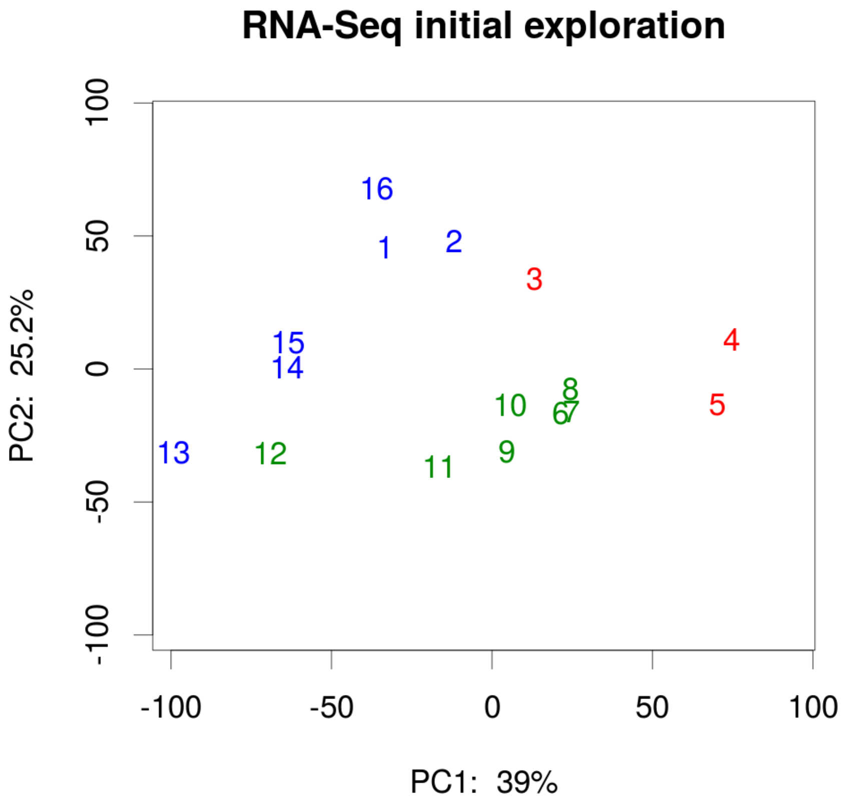


Figure 17. Score plot for the principal components 1 and 3, obtained with the RNA-Seq data of 5992 genes. In green, the time points associated to the RB phase, in blue, the time points associated to the RC phase, and in red, the time points associated to the OX phase.

3.4.2 Differential expression analysis

As detailed in Section 2.4.2, maSigPro software was used to perform a differential expression analysis and clustering of the genes, based on the RNA-Seq data. Thus, a model per gene was generated where the response variable was the gene expression and time was the predictor.

In order to decide the best polynomial degree to model the RNA-Seq data, 7 different degrees were tested and a different number of differentially expressed genes (DEGs) was obtained in each case. In addition, PCA was applied on the data corresponding to each of the 7 groups of DEGs. Table 9 shows the number of DEGs and the percentage of variance explained by the two principal components of the PCA for each of the 7 models. The models with degree 2 and 3 showed the highest percentages of explained variance and observing the PCA score plots, the best separation of the time points was achieved also for these models.

Table 9. Number (Nr) of differentially expressed genes (DEGs) for each of the polynomial degrees, and percentage of explained variances with the first principal components associated to the PCA with the DEGs.

Degree	Nr of DEGs	Explained variance with PC1 and PC2
2	2021	78.48%
3	3397	77.86%
4	4774	69.0%
5	4730	68.13%
6	4315	69.02%
7	5053	68.34%
8	4650	69.10%

Thus, this first exploration seemed to point to degree 2 or 3. To confirm this choice, we also analysed the adjusted coefficients of determination (R^2) values of the models. First, the R^2 value of the regression model for each of the DEGs was retrieved for the 7 tested degrees. Next, the R^2 values were adjusted so the regression models with different polynomial degrees could be compared (Figure 18). The largest increase of the adjusted R^2 is observed from degree 2 to 3, and a smaller improvement can be noticed at degree 4. For higher degrees, the differences are insignificant.

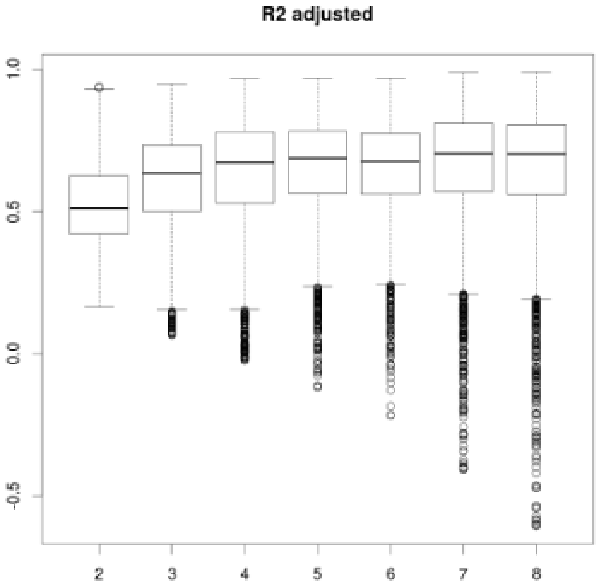


Figure 18. Adjusted R^2 values boxplot for the differentially expressed gene models, regarding the 7 different degrees tested.

Taking into account the principle that: (i) a simpler model is generally better when comparing models with similar performance (because a more complex one can tend to model noisy or stranger patterns or to overfit the data); (ii) the subset of DEGs obtained with degrees 2 or 3 were the ones showing a better separation of the observations in the PCA analysis; and (iii) beyond a polynomial degree of 3 or 4 there is not a notorious increase of the adjusted R^2 values of the regression models obtained for the DEGs; a degree of 3 was chosen to model the behaviour of the RNA-Seq data.

Lastly, for the DEGs obtained with a polynomial degree of 3 (3397), only those whose associated R^2 value was higher than 0.6 were selected to continue with the analysis.

Based on this criteria, 2552 genes were kept. The reason behind this R^2 filter was keeping only those genes for which most of its variance, at least 60 per cent, could be explained by the model. These genes were supposed to be associated to the Yeast Metabolic Cycle, so their expression levels should change in a way that time could explain most of their variation.

Figure 19 shows the scores from the final PCA applied on the 2552 genes selected and on the 3929 DEGs considered in (Kuang, et al., 2014). A better representation of the distribution of the time points along the YMC can be observed with the 2552 selected genes in this study against the 3929 previously considered. These PCA results support the differential expression strategy followed.

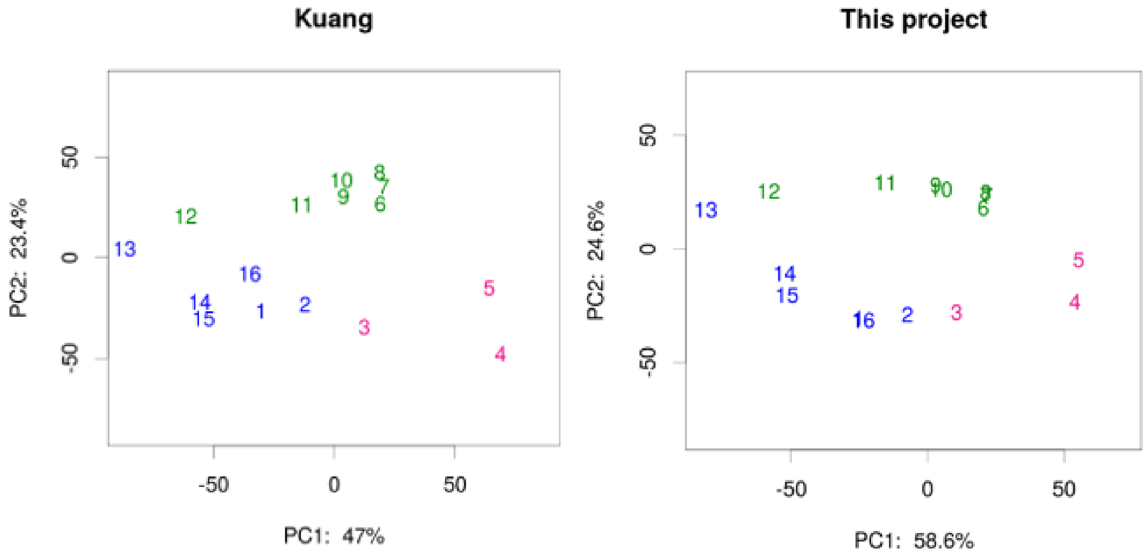


Figure 19. Score plot for the PCA obtained with the 3929 DEGs in Kuang et. al (left) and with the 2552 DEGs selected in this project (right). In green, time points associated to the RB phase, in blue, time points associated to the RC phase, and in red, time points associated to the OX phase.

The PCA on our 2552 selected DEGs (Figure 19, right) shows how the time points are separated regarding the metabolic phase in which they were obtained. Moreover, they are ordered from 1 to 16 representing nicely the distribution of the time points along the YMC. It can be observed, for instance, how the time points 1 and 16 are really close, as expected, since they represent the beginning and the end of the cycle (Figure 1). Another example of the meaningfulness of this representation is that time point 12 is more similar to time point 13from RC phase than to time point 11 from its same phase (RB). Based on the levels of oxygen consumption (Figure 1), this makes sense, since the levels of oxygen are more similar between time points 12 and 13 than between time points 11 and 12.

3.4.3 RNA-Seq genes clustering

Next step was clustering the 2552 DEGs according to their temporal profiles. This analysis was carried out with maSigPro software, that applies k-means algorithm. We set the number of clusters k to 3, which corresponds with the 3 main metabolic phases in the YMC: OX, RB and RC (Kuang, et al., 2014).

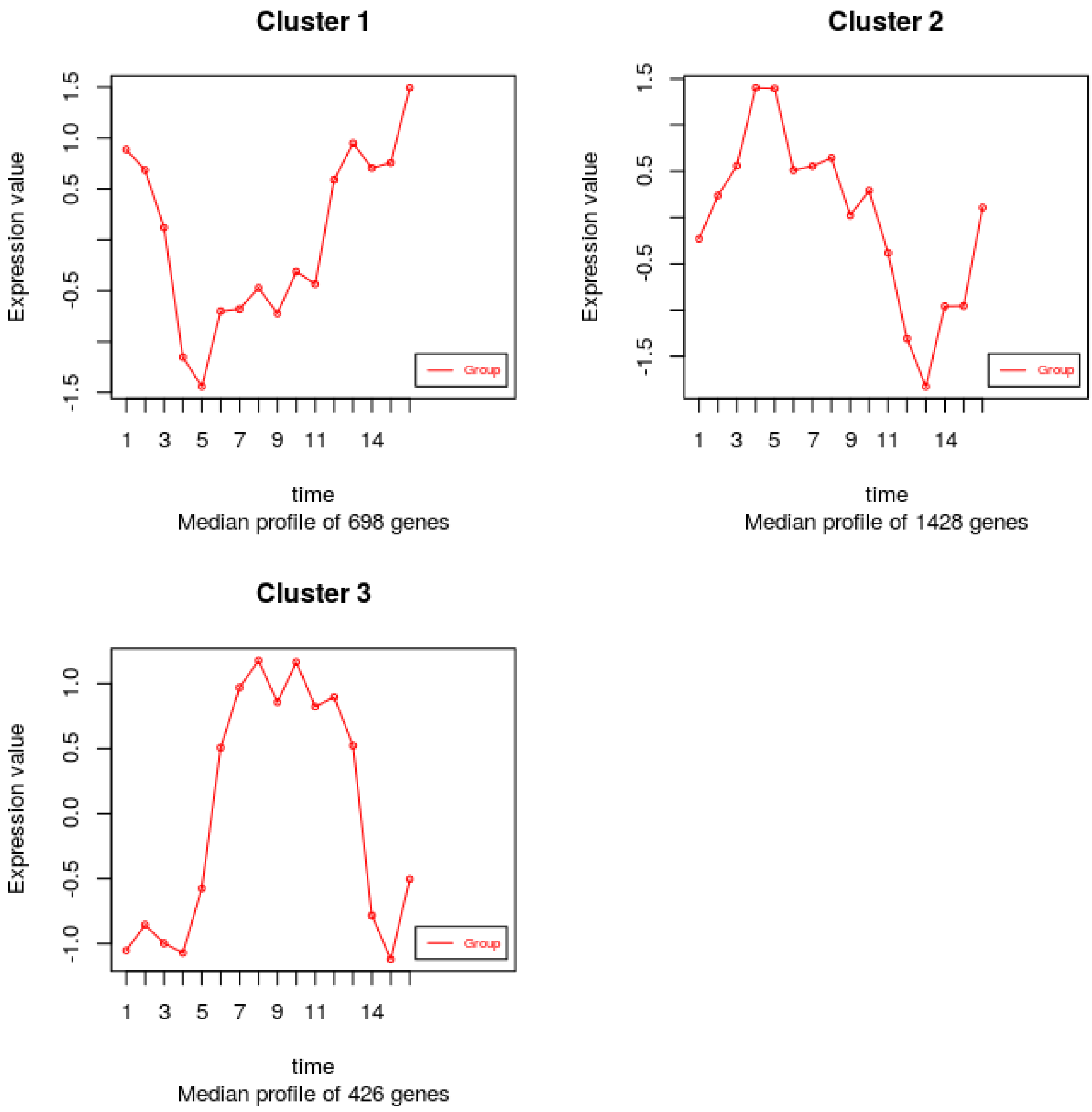


Figure 20. RNA-Seq clustering of the 2552 selected genes using K-means specifying 3 clusters. The mean expression level, for each time point and for all the genes in each cluster is represented. Horizontal axis corresponds to the time points, from 1 to 16, and the vertical axis corresponds to the expression levels.

The number of genes assigned to each of the three clusters was 1428, 698 and 426 (Figure 20). Focusing the attention on the mean profile for the different clusters: the cluster 3 was associated with the metabolic phase RB since the higer expression values of the genes present on it were given approximately between the time points 6 and 12, the ones located in the RB phase. Based on this reasoning, cluster 2 was associated with the RC metabolic phase and the cluster 1 with the OX phase. Thus, from the 2552 genes, 1428 were associated to the OX metabolic phase, 698 to the RC metabolic phase and 426 to the RB metabolic phase.

(Kuang, et al., 2014) also clustered their 3929 DEGs into 3 clusters with 946 in the RB phase, 1441 in the RC phase, and 1542 in the OX phase. We compared the quality of the clustering of this study to the quality of our clustering.

Venn diagrams in Figure 21 and Table 10 compare the number of genes per cluster in both studies. More than 80% of the genes of each cluster in this project were also included in the same cluster in (Kuang, et al., 2014). Thus, in general terms, the clusters from this project, despite being smaller, were quite similar to the ones from Kuang et al. However, there were some differences. In order to assess the clustering quality of each study, one silhouette plot per clustering was done (Figure 22 and Figure 23).

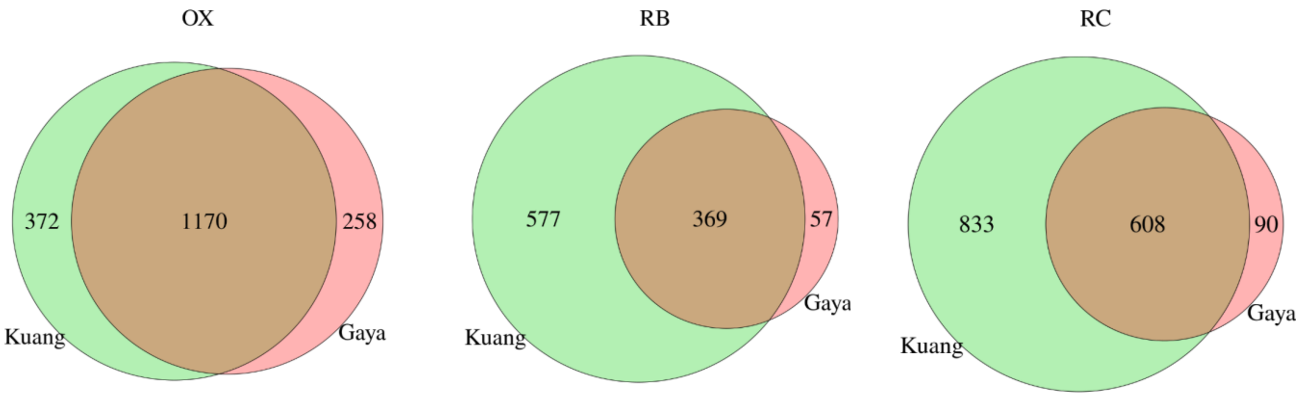


Figure 21. Comparison of the clustered genes in this project (Gaya), and in (Kuang, et al., 2014). From left to right, OX, RB and RC phase Venn diagrams. In green, genes clustered only in (Kuang, et al., 2014). In red, genes only clustered in this project. In brown, genes in common for both studies.

Table 10. Comparison of clustering results in this project and in (Kuang et al., 2014). First column: Number of clustered genes in this project, also included in the same cluster in (Kuang, et al., 2014). Second column: Number of clustered genes in this project not included in Kuang's clusters. Third column: Percentage of genes clustered in this project that are also included in Kuang's clusters.

Cluster	In Kuang cluster	Not in Kuang cluster	% included in Kuang cluster
RB	369	57	86.6
RC	608	90	87.1
OX	1170	258	81.9

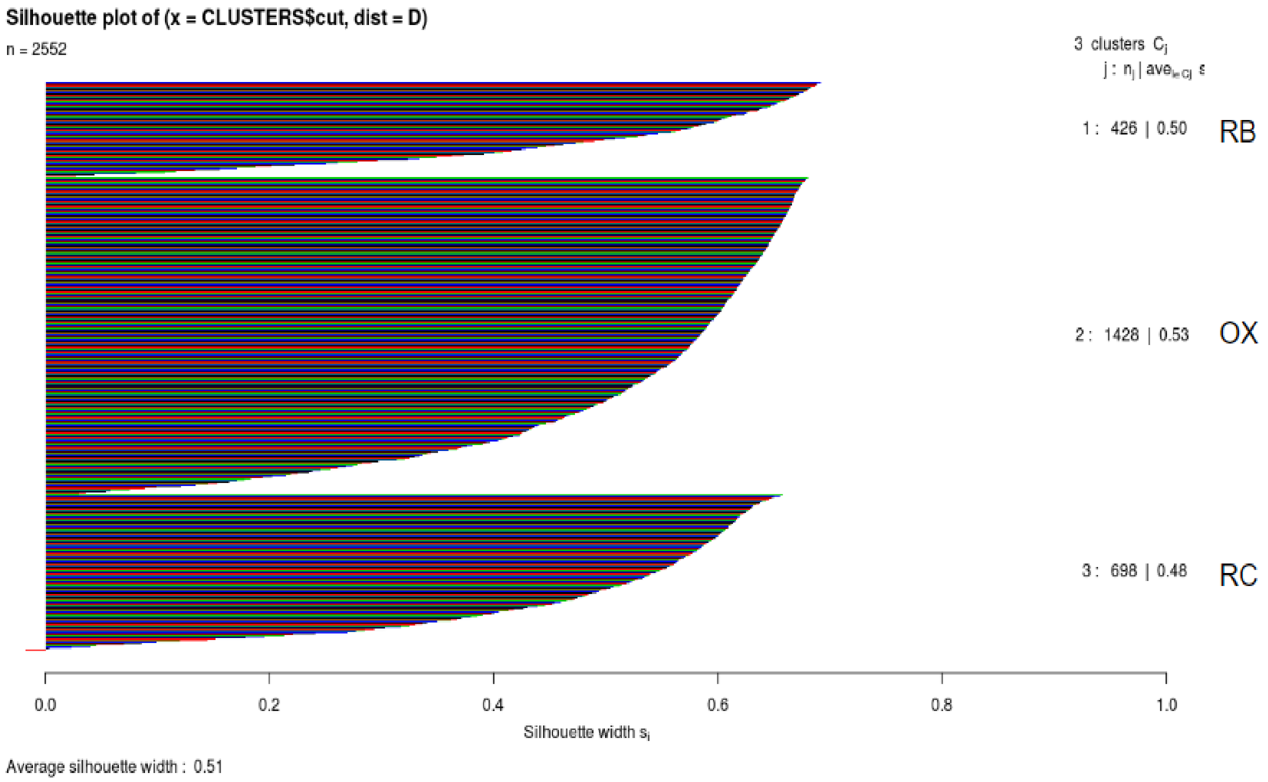


Figure 22. Silhouette plot of the gene clustering in this project.

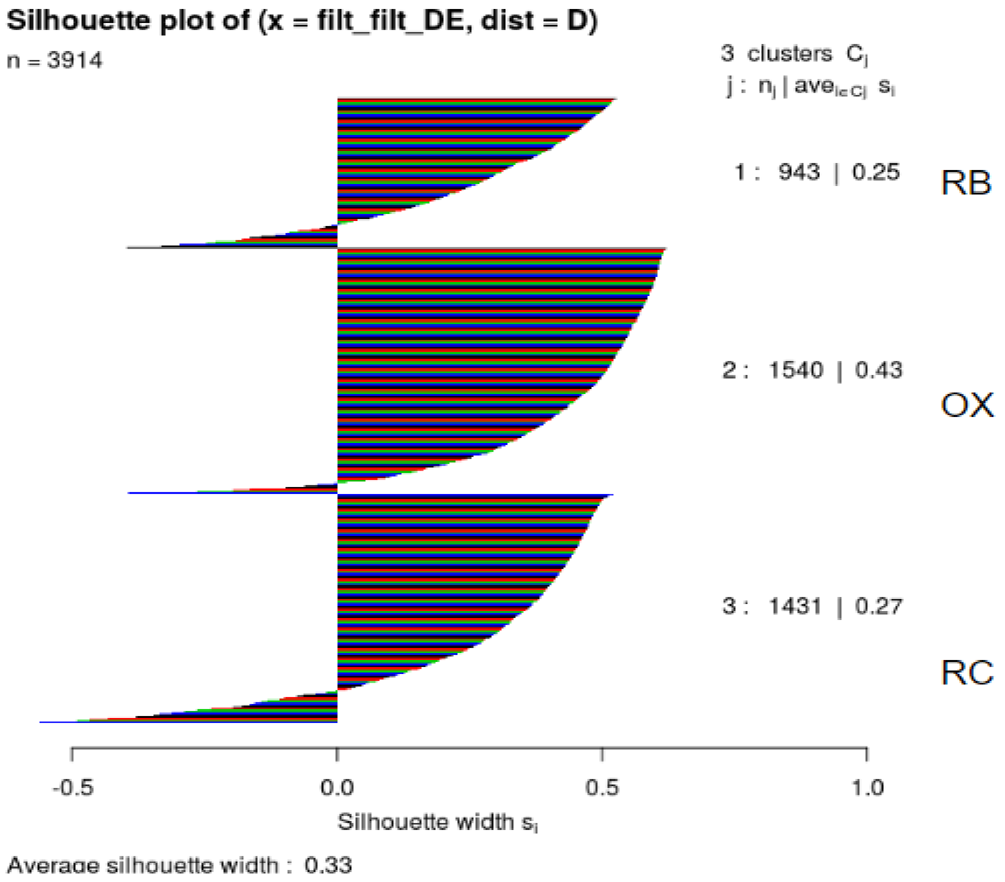


Figure 23. Silhouette plot of the gene clustering in **(Kuang, et al., 2014)**.

For Kuang’s clustering (Figure 23), not all of the 3929 DEGs considered in their study were evaluated, having instead just 3914, since we observed in their results that some genes were simultaneously assigned to more than one cluster, which did not make sense because they performed a K-means clustering. Moreover, some other genes had been considered as differentially expressed and therefore clustered, despite their RNA-Seq levels for all the time points were 0. These discrepancies were probably due to the fact that, in Kuang’s paper, RNA-Seq data results were compared with the ones from a similar project performed with microarrays (Tu, et al., 2005), and possibly both data sets were somehow merged. Thus, we excluded these problematic ones for the Silhouette analysis.

The average silhouette width for all the clusters in this project is about 0.5, while the one from (Kuang, et al., 2014) is about 0.3. In the clustering of this project, just one gene can be considered as wrongly clustered (the one at the bottom of the plot, presenting a negative value), while in Kuang’s clustering many observations seem to be wrongly classified, based on the negative silhouette distances.

These results justify why we decided to repeat part of the analysis that was done in (Kuang, et al., 2014), to be sure that we had a robust selection of DEGs and a more trustable clustering, since these two results will be key to proceed with the omics integration analysis.

3.5 Omics integration

3.5.1 N-PLS

N-PLS method was used as a first exploratory approach to integrate the gene expression and chromatin status data and analyse the relevance of the different histone modifications in the global regulation of gene expression in the YMC.

Regarding the assignment of the different variables to the different modes, genes were placed in the first mode, time points in the second mode and the different histone modifications in the third mode (Figure 3).

The analysis of the core matrix showed that the most important core entry was (2,2,2), which was explaining 60.7% of the variance in the data, and the second most important was (2,2,1), which explained 23.9% (Table 11). Thus, more than 80% of the variance was explained with the usage of these two entries. As an example of the interpretation of the core entries, the entry (2,2,1) means that 23.9% of all the variance in the data being analysed (RNA-Seq and ChIP-Seq) is explained with the second component of mode 1 (genes), the second component of mode 2 (time points) and the first component of mode 3 (histone modifications). Figure 24, Figure 25 and Figure 26 show the loading plots for all the modes, attending to the RNA-Seq and ChIP-Seq data.

Table 11. Elements of the core sorted by the percentage of explained variation.

	Core element	Explained variation of the core
1	(2,2,2)	60.69%
2	(2,2,1)	23.92%
3	(1,1,2)	5.04%
4	(1,1,1)	4.64%
5	(1,2,1)	2.69%
6	(1,2,2)	2.47%
7	(2,1,2)	0. 34%
8	(2,1,1)	0.21%

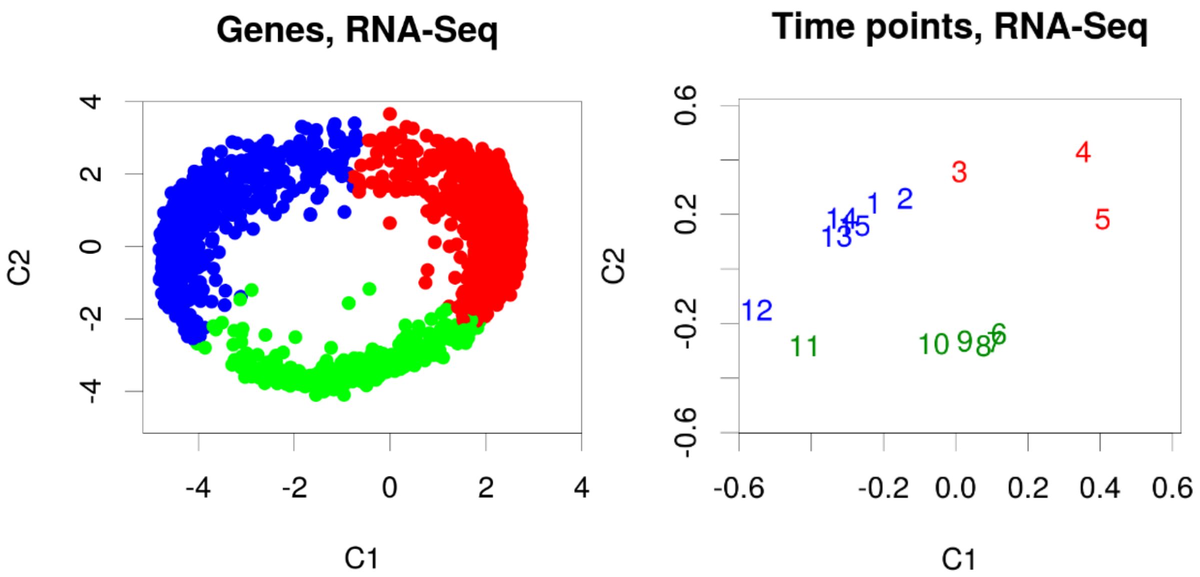


Figure 24. N-PLS loading plot for RNA-Seq data. Components 1 and 2 are displayed for modes 1 (genes, left), and 2 (time points, right). In green, either genes or time points associated to the RB phase. In blue, either genes or time points associated to the RC phase. In red, either genes or time points associated to the OX phase.

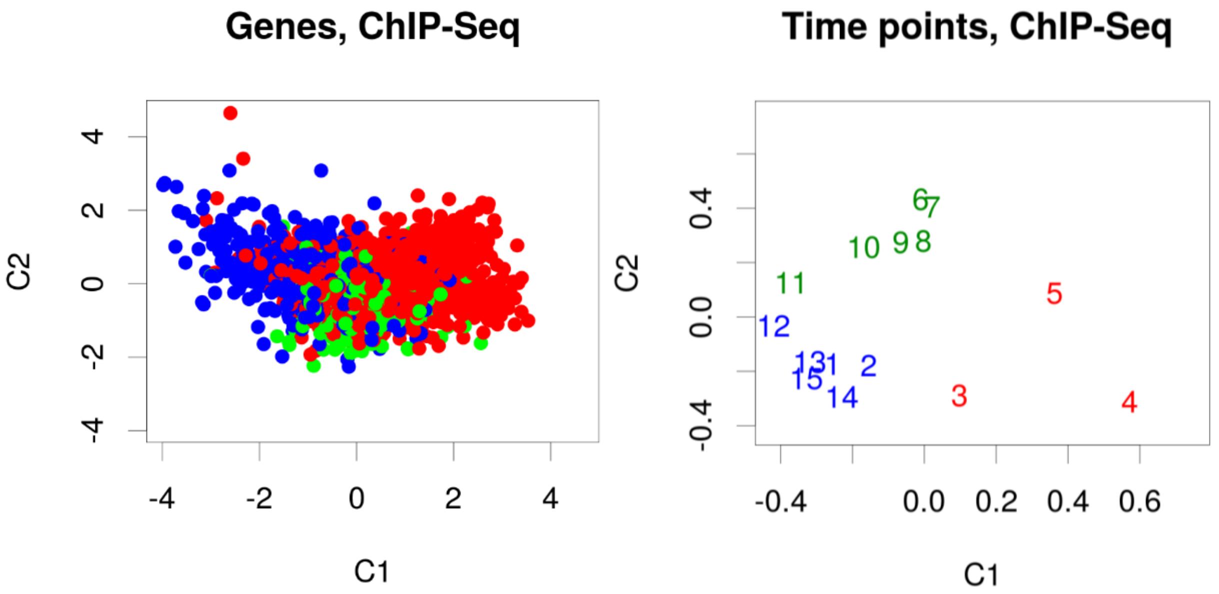


Figure 25. N-PLS loading plot for ChIP-Seq data. Components 1 and 2 are displayed for modes 1 (genes, left), and 2 (time points, right). In green, either genes or time points associated to the RB phase. In blue,

either genes or time points associated to the RC phase. In red, either genes or time points associated to the OX phase.

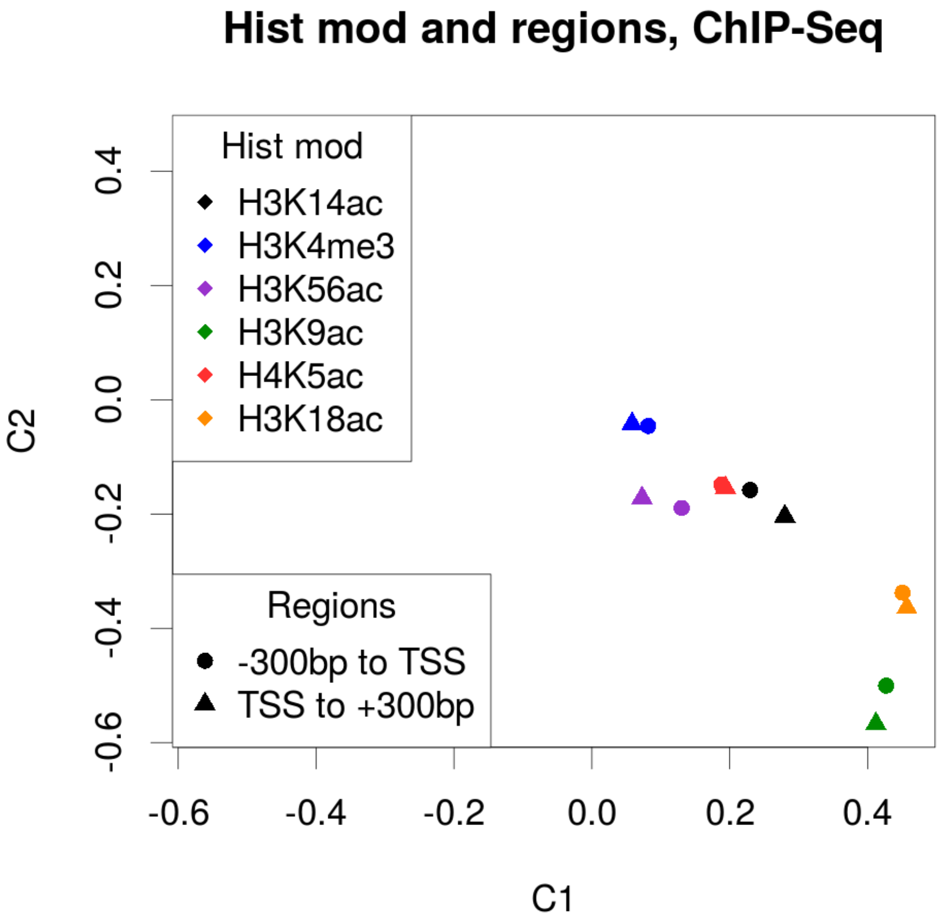


Figure 26. N-PLS loading plot for ChIP-Seq data. Components 1 and 2 are displayed for mode 3 (histone modifications and genomic regions). For each histone modification there are two regions: circle shape is for -300 bp to TSS regions, triangle shape is for TSS to +300 bp regions.

Based on the two most relevant entries of the core matrix ((2,2,2) and (2,2,1)), and first analysing the third mode (histone modifications), both components 1 and 2 can be considered as notoriously relevant for the explanation of the data variance. In absolute value, H3K9ac and H3K18ac are the histone modifications presenting the highest loadings, H3K4me3 the histone modification presenting the smallest loadings and H3K14c, H3K56ac and H4K5ac present intermediate loadings in comparison with H3K4me3 and the histones modifications with the highest loadings (H3K9ac and H3K18ac) (Figure 26). The two different regions considered (-300bp to TSS and TSS to +300 bp) for each histone modification are close to each other. The second component of the second mode (Figure 24 and Figure 25, right plot) is separating mainly those time points associated with the metabolic state of the cells when more oxygen is consumed along the YMC (time points 6, 7, 8, 9, 10), against the rest of time points. Finally, the second component of the first mode (Figure 24 left) shows how the genes associated to the RB (green) phase are separated from the genes associated to the OX (red) and RC (blue) phases. Figure 25 shows the same trend, that is, the second component aims to separate the RB genes from the RC and OX genes. However, this separation is not as clear as in RNA-Seq data.

3.5.2 MORE

3.5.2.1 MORE with histone modifications as predictors

The aim of this approach was to find which histone modifications significantly regulate the expression levels of each gene individually, so a MORE regression model was obtained for each gene to relate the gene expression values (response) to the chromatin status given by the different histone modification quantification values for that gene (predictors) and significant histone modifications were selected for each gene at a significance level of 5%.

As indicated in (Berger, 2007), all the histone modifications to be studied in this project after the pre-processing steps (H3K14ac, H3K4me3, H3K56ac, H4K5ac, H3K9ac and H3K18ac) have a role on transcriptional activation. As a consequence, negative correlations between gene expression and histone modification are difficult to interpret and likely to be spurious, so it was decided to filter out of MORE results those significant histone modifications with negative regression coefficients in the final regression model for each gene. Figure 27 and Figure 28 show the number of genes significantly regulated by each histone modification, globally and per cluster, respectively.

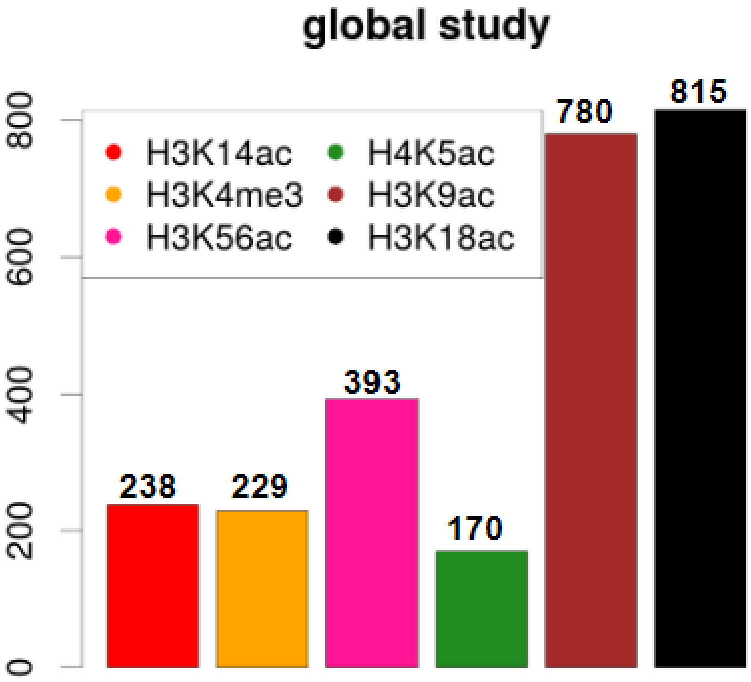
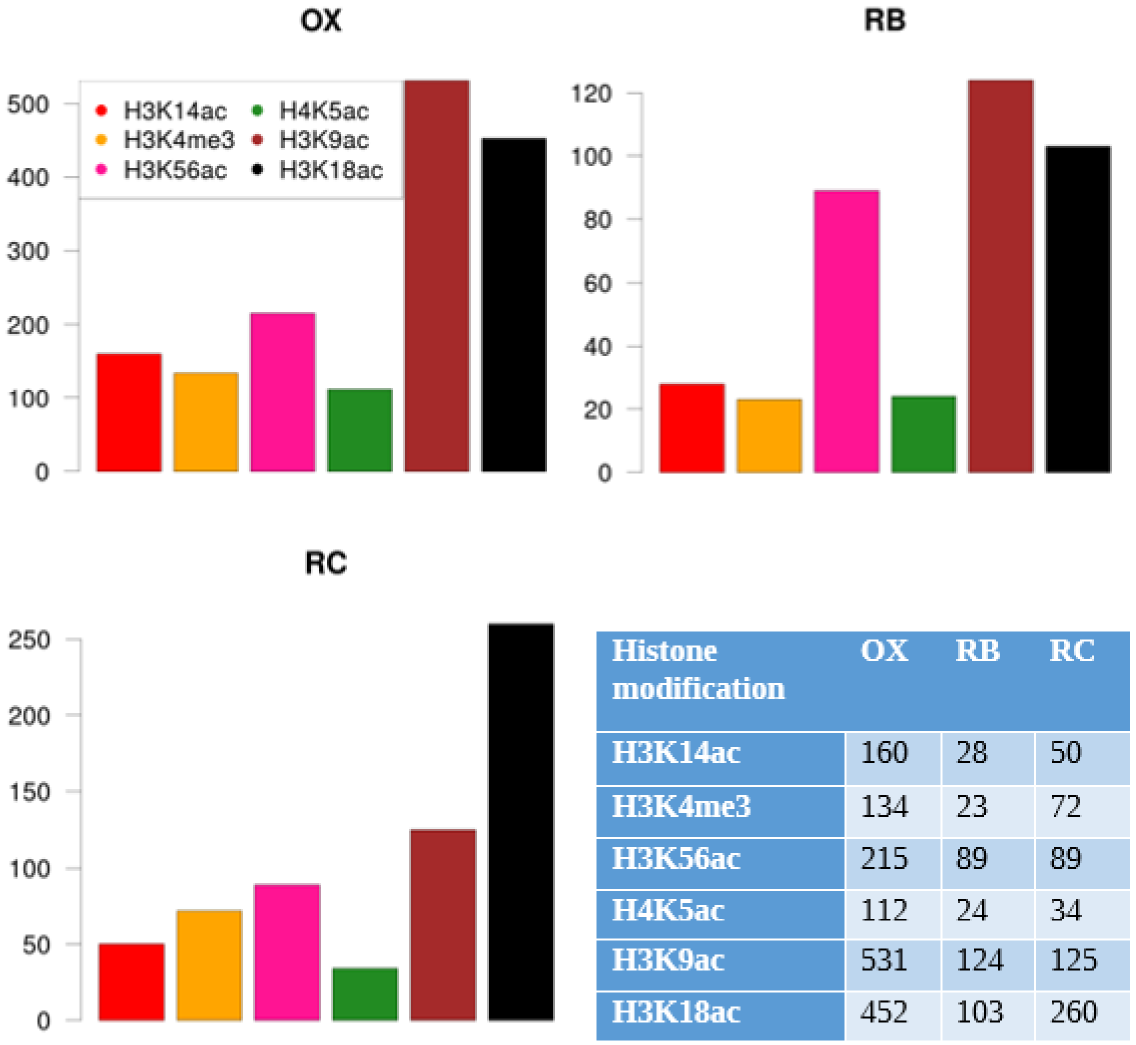


Figure 27. MORE results. Number of significantly associated genes to each histone modification (p-value < 0.05).



Histone modification	OX	RB	RC
H3K14ac	160	28	50
H3K4me3	134	23	72
H3K56ac	215	89	89
H4K5ac	112	24	34
H3K9ac	531	124	125
H3K18ac	452	103	260

Figure 28. MORE results per gene cluster. Number of genes significantly associated to each histone modification per cluster. Bottom right table provides the same information as the plots, the number of genes significantly associated by MORE to each histone modification, per cluster.

These results indicate that H3K9ac and H3K18ac, either globally or per cluster, are the histone modifications which regulate a higher amount of genes, confirming the N-PLS results. Regarding the RC phase, apparently the role of H3K18ac is higher in comparison to H3K9ac in the regulation of the gene expression, since it is regulating a considerably higher number of genes (260 vs 125). H3K56ac regulates a notorious bigger proportion of genes in the RB phase compared with the OX and RC phase.

3.5.2.2 MORE with transcription factors as predictors

The study of transcription factors was included in the project as an additional layer for the comprehension of the regulation of the YMC. According to Yeastract (Teixeira, et al., 2018), 109 transcription factors (TFs) were included among our DEGs. The different target genes for each TF were also retrieved from Yeastract.

Similarly to the previous section, MORE was again applied to model gene expression as a function of transcription factor expression, in this case. The results indicated that 105 TFs were significantly regulating 2480 genes. To facilitate the interpretations of these results, we selected the significant TFs that were regulating a significantly higher proportion of genes after MORE than before MORE (Yeastract database) and named them as relevant TFs (see Section 2.4.6). We found 13 relevant TFs (see gene names, adjusted p-values and odds ratio in Table 12).

Table 12. Selection of relevant TFs from MORE results. Gene name, adjusted p-value and odds ratio for the significant TFs in the Fisher's Exact test.

	Adjusted p-value	Odds ratio
YOR028C	2.60e-04	1.38
YLR403W	1.23e-16	1.60
YHR084W	1.33e-03	1.26
YPL254W	1.06e-13	2.01
YPL177C	4.50e-03	1.40
YOR363C	1.05e-09	1.79
YIL101C	8.76e-05	1.76
YGL209W	1.09e-12	3.64
YDR451C	4.86e-10	1.83
YLR278C	4.23e-03	1.67
YIL130W	1.82e-04	2.39
YLR014C	3.86e-04	2.65
YML113W	7.04e-03	2.05

Due to time constraints, not all the relevant TFs in Table 12 were studied, but only those with an adjusted p-value < 1e-08. Hence, YLR403W, YPL254W, YOR363C, YGL209W and YDR451C were selected to continue the study with them.

3.6 Functional Enrichment Analysis

3.6.1 Histone modifications

Concerning the genes significantly associated by MORE method to each histone modification per cluster (Figure 28), different functional enrichment analyses were done. The most relevant results on shared and unique traits of each phase are detailed below.

OX phase

OX phase shows a common trend of regulating translation, ribosomal machinery and nucleotides metabolism. Except H4K5ac, every other histone modification was involved in aminoacid metabolism; being H3K18ac, H3K14ac and H3K4me3 specifically linked with one-carbon metabolism and methylation. H3K56ac and H4K5ac showed overrepresentation of helicase activity within the enriched functionalities, being this latter histone modification the only one presenting enhanced cell cycle regulation functionalities.

RB phase

RB phase is characterized by a regulation of sugar metabolism by all the histone modifications, coupled with a regulation of mitochondrial activity, from which H3k4me3 is excluded. Genes regulating lipid metabolism are coordinated by H3K9ac, H3K18ac, H3K14ac and H3K56ac, whereas aminoacid metabolism is enriched in genes regulated by H3K14ac, H3K56ac, H4K5ac and H3K4me3. H3K9ac and H3K56ac are linked to ribosomal activity. Cell cycle appears to be regulated by H3K18ac, H3K14ac and H3K4me3; H3K9ac, H3K14ac, H3K56ac and H4K5ac are coordinating secondary metabolism, some of these being involved in antibiotic biosynthesis. H3K4me3 appears to be the only modification to be significantly regulating ethanol biosynthesis.

RC phase

This phase is arguably the most diversified among the histone modification regulatory functions. Carbon metabolism and Oxidation-reduction processes are regulated by all modifications except H3K9ac, which appears to coordinate ethanol metabolism with H3K14ac and H3K4me3; and fatty acid metabolism together with H3K18ac, H3K56ac and H3K4me3. H3K18ac and H3K56ac combine to target genes with amino-acid metabolism and cell division functionalities. Oxidative phosphorylation appeared to be marked by all modifications but H3K56ac, which seemed to be the only one related to genes involved in histone acetylation.

3.6.2 Transcription factors

YLR403W appeared involved in the transcription of ribosomal proteins, nutrient response, G2/M transitions, DNA damage, and histone exchange. YPL254 presented a role in translation, aminoacid and nucleotides biosynthesis and nucleotide cleavage. YOR363C was associated to B-oxidation, peroxisome organization, oleate-driven transport, glucose starvation response and the ribosome and glyoxylate cycle. YGL209W was contributing to glucose-induced gene expression and mitochondrial fusion, besides being involved in endonuclease cleavage, ribosome, transferase, kinase activity and carbon transport. Last, for YDR451C, the functions related to ribosomes, methylation, negative response of invasive growth and cell cycle regulatory control were enriched.

3.7 Significant association results

Significant association analysis (Section 2.4.7) was applied to study if a given pair TF-histone modification was significantly co-regulating a set of genes. Two of all the tested pairs resulted significant, which corresponded to the histone modification H3K18ac with the TFs YPL254W and YOR363C, with adjusted p-values 0.0076 and 0.017, respectively.

3.8 Regulatory network

A regulatory network was created in order to depict the relevance of the different histone modifications and transcription factors on driving gene expression changes in the context of the YMC. The regulatory network was created with Cytoscape (Shannon, et al., 2003) and the result is shown in Figure 29. This figure shows that H3K9ac and H3K18ac are the histone modifications regulating a higher number of genes (bigger nodes), while H4K5ac is regulating the smallest amount of genes. It can be seen also that YLR403W and YGL209W are the TFs regulating the highest and lowest number of genes, respectively. Regarding the proportion of regulated genes per metabolic phase (pie charts) in the context of the histone modifications, the proportion of regulated genes per metabolic phase is approximately the same for all the histone modifications, whereas in the context of the TFs, some differences are noticeable. For instance, YLR403W regulates a higher proportion of genes in the OX phase than in the RB and RC phases; and YDR451C regulates a higher proportion of genes in the RB phase than in the OX and RC phases.

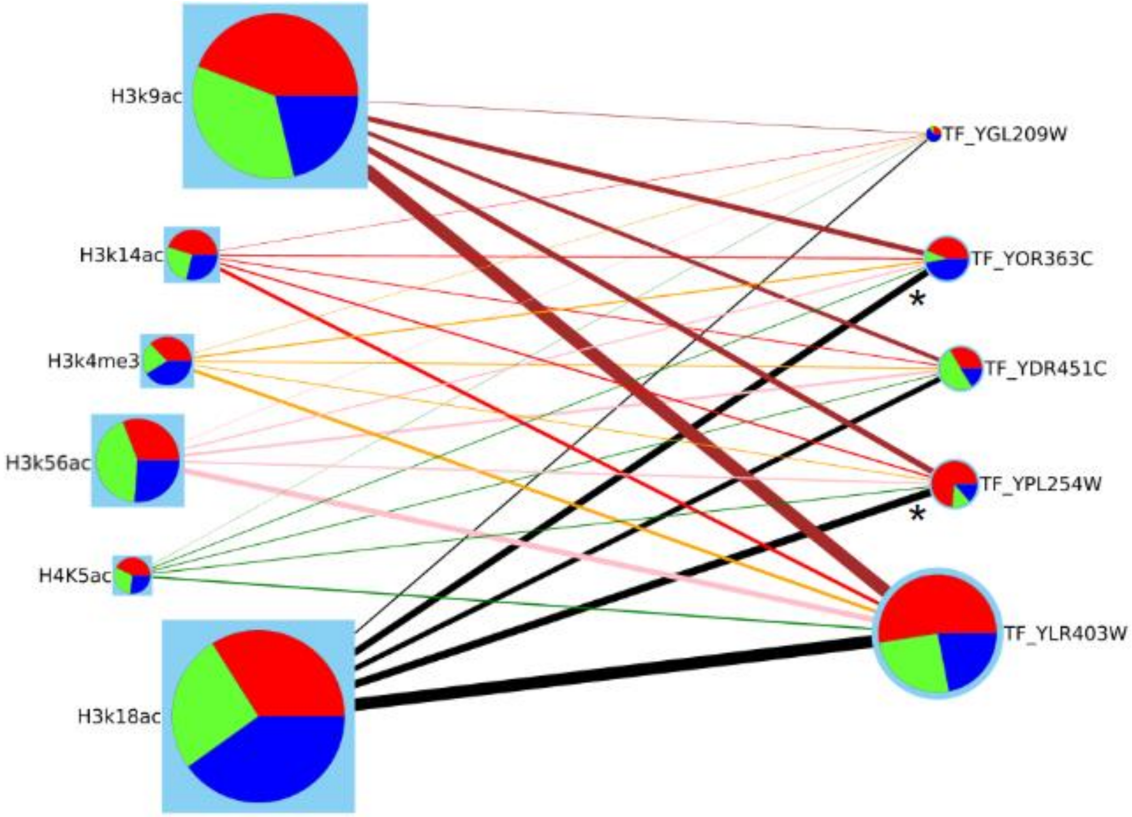


Figure 29. YMC Regulatory network. The shape of the node depends on the type of regulator: transcription factors are represented with a circle and histone modifications with a square. The size of the node is determined by the number of genes significantly regulated by the regulator. The thickness of the edges determines the number of genes simultaneously regulated by the connected nodes (regulators). An asterisk on an edge indicates a significant association between the nodes connected by that edge. The construction of the pie chart was done as follows: 1- From the subset of genes regulated by the regulator of interest, the number of genes associated to each cluster was calculated. 2- the proportion of those number of genes over their respective clusters was obtained. 3- with that proportions the pie chart was elaborated. For instance, if the 3 slices present the same size it means that the studied regulator controls the expression of the same proportion of genes in the 3 clusters. The colours of the clusters are: OX (red) RB (green) and RC (blue).

4. Discussion

This project studies the regulatory relationship that histone modifications may have over gene expression in the YMC. The differential gene expression analysis was performed to focus the regulatory study (histone modifications and transcription factors) on the genes changing in the YMC, because the rest of genes were not relevant for this study. This set of DEGs provided a more solid landscape to draw relevant conclusions from the omics integration analysis.

The DEG clusters proposed here were very similar to the ones in (Kuang, et al., 2014), with more than 80% of overlap. However, the differences in the total number of DEGs considered, 2552 in this project and 3929 in (Kuang, et al., 2014), and in the quality of the clustering revealed that many genes were wrongly classified in Kuang's study and that the new clusters of this master thesis were more reliable and could lead to more biologically meaningful results. The reason of the differences between both strategies may be that in (Kuang, et al., 2014) the differential expression analysis was not done considering the time variable to model gene expression as it was done here with maSigPro. Moreover, the additional R^2 filter we applied could have helped to select genes with a well-defined temporal profile and discard the noisy ones.

Pre-processing again the ChIP-seq samples in order to properly quantify the chromatin status also allowed us to have more trustable data to perform the integration analysis, since the complexity of the integration strategies require even more clear signals than omic-independent analysis.

The multi-omics exploratory analysis of this work consisted in applying a multiway dimension reduction method, N-PLS, which is an extension of the multivariate PLS methodology for data structures with more than two dimensions. In this case, the response variable was gene expression data and it was bidimensional but the predictor matrix collecting all the histone modification experiments was three-way. The N-PLS technique allowed us to explore the data from a global perspective, to confirm that the spatial conformation of genes and time points was in concordance with the YMC, and to get a first impression of the histone modifications that were more prone to be regulating gene expression as well as the relationship among them.

In the first place, the N-PLS results showed that the genes clustering for histone modifications was not as clear as the separation based on gene expression data. A possible reason for this could be that many of the differentially expressed genes are not strongly regulated by histone modifications. As a consequence, their associated histone modification levels would not be relevant enough to be clustered according to the metabolic phase. Another reason is that the ChIP-Seq measurements are noisier than RNA-Seq data, so it is logical to get lower resolution.

However, the N-PLS analysis grouped the two regions defined for each histone modifications (-300bp to TSS and TSS to +300 bp), which indicated that the effect of the studied histone modifications was similar upstream and downstream the TSS. This suggests that in future works, a unique region extending from -300 bp to +300 bp with respect to the TSS should be defined for these histone modifications. Another conclusion from these plots was that histone modification H3K18ac, that was the only one provided by our collaborator instead of downloaded from GEO repository, had similar behaviour to the rest of histone modifications (it was not an outlier), which is a good result because otherwise it could not have been included in the study since it was not possible to correct the effect of this batch.

The N-PLS method also confirmed that time points associated to maximum oxygen consumption or low oxygen consumption clustered together, as well as the different phases, and revealed that H3K18ac and H3K9ac were the most relevant histone modifications involved in the regulation of the gene expression changes in the YMC, which coincided with the MORE results that will be discussed next. In brief, N-PLS served to check the quality of the analysed data and to highlight that histone modifications play a relevant role in the switch of the metabolism from low to high oxygen consumption state. In addition, the model indicated that the most important differences regarding the histone modification levels of the differentially expressed genes in the YMC were associated to such states of oxygen consumption.

The second integrative approach used in this project was MORE method that models gene expression as a function of different potential regulators such as histone modifications or transcriptions factors by applying Generalized Linear Models together with different variable selection strategies. Thus, MORE models allowed for finding those histone modifications or transcription factors significantly regulating each specific gene during YMC.

MORE results for histone modifications as predictors confirmed the N-PLS results that H3K18ac and H3K9ac had a higher effect on gene expression regulation than the rest of studied histone modifications. This was concluded from the fact that H3K18ac and H3K9ac were significantly regulating a larger amount of genes, both globally and per gene cluster. While apparently having the same relevance in OX and RB phases, H3K18ac seems to be more important along the RC phase, since it was regulating more than twice the amount of genes regulated by H3K9ac.

The functional enrichment (FE) analyses of genes significantly regulated by H3K9ac and H3K18ac revealed that these genes were significantly involved in processes related to each of the three metabolic phases, which was expected since they regulate a large amount of genes in each cluster. Thus, H3K9ac and H3K18ac were found to be responsible of the regulation of the main functionalities associated to each metabolic phase, which are, principally, the synthesis of ribosomes and amino acids in OX; the regulation of mitochondrial genes and the activation of metabolic pathways for carbon degradation in RB; and the degradation of fatty acids in RC; these observations were in line with the results described in (Kuang, et al., 2014). However, no relevant functional terms were enriched for H4K5ac, probably due to the small number of genes it regulates which makes it to be considered as the least important of the different studied histone modifications in the regulation of the YMC. H3K4me3 is significantly involved in the synthesis of amino acids along all the phases, specifically related with the cycle of the one-carbon metabolism. This is interesting since it has been described in (Mentch, et al., 2015) how the histone methylation dynamics (H3K4me3 levels in particular) and the regulation of the gene expression occur through the one-carbon-metabolism. Hence this project gives more support to their results relating the metabolome with the chromatin status and the implications over gene expression. Different studies (Cai, et al., 2011), (Wellen, et al., 2009) have also described how the levels of different metabolites, acetyl-

CoA and ATP-citrate, have an impact over the histone acetylation levels of the DNA. Acetyl-CoA is derived from glucose, and the role of ATP-citrate in acetylation, as detailed in (Wellen, et al., 2009), is due to the conversion of glucose-derived citrate into acetyl-CoA. Thus, the processing of the glucose has a direct impact on the acetylation levels of the chromatin. In (Zhao, et al., 2010) more information is given combining metabolism and chromatin status, and shows how the concentration of metabolic fuels, such as glucose, amino acids, and fatty acids influence the acetylation status of different metabolic enzymes. The proteins lysine acetylation is described as a key posttranslational modification in cellular regulation, in particular through the modification of histones and nuclear transcription regulators. Taking into account that the YMC is given under glucose limited conditions, which is directly going to cause that the available levels of glucose shift along time, and based on the exposed references, it is expected to have oscillations in the concentration of this compound which will derive on the dynamic acetylation of the DNA. Therefore, having H3K9ac and H3K18ac as the most relevant histone modifications regulating the gene expression of the YMC was expected, and reinforces what had been previously reported.

Since both N-PLS and MORE results showed that many genes were not being regulated by histone modifications across YMC, it was decided to analyse the role of differentially expressed transcription factors (TFs) on gene expression regulation, with the goal of adding new information which could help to better understand the regulatory mechanisms behind the YMC. To do that, we again applied MORE but taking TFs as predictors. An enrichment analysis on these results revealed the most relevant TFs in terms of proportion of significant regulations.

The results for the FE analysis on the target genes of relevant TFs were compared with previously published results. YLR403W, also named as Sfp1, was found to be regulating the transcription of ribosomal proteins, nutrient response, G2/M transitions and DNA damage in concordance with (Marion, et al., 2004), (Xu & Norris, 1998). In (Marion, et al., 2004), Sfp1 is defined as a stress and nutrient-sensitive regulator. In addition, a clear role of Sfp1 in histone exchange is found among the results of this project. Attending to YOR363C, also named as PIP2, there was also evidence in the literature regarding its role in B-oxidation, peroxisome organization and oleate-driven transport (Baumgartner, et al., 1999) (Rottensteiner, et al., 1996), as we also found in our results. In addition, we also found a role in the ribosome and glyoxylate cycle and a response effect to glucose starvation. With respect to YGL209W, whose standard name is MIG2, the literature shows that the levels of glucose regulate the distribution of MIG2 (Fernández-Cid, et al., 2012) and that the filamentous growth of MAPK pathway responds to glucose starvation through MIG2 and MIG1 (Karunanithi & Cullen, 2012). Our results confirmed its role in glucose-induced gene expression and, additionally, new roles in endonuclease cleavage and kinase activity among others were found. Analysing the results of YDR451C, YHP1 as standard name, the literature mentions its role in the regulation of the cell cycle (Kunoh, et al., 2000), (Pramila, et al., 2002). In this project, this TF was found to be involved in functions related to ribosomes, methylation and negative response of invasive growth. Lastly, YPL254W was considered in the literature as a relevant protein adaptor for the structural integrity of the SAGA complex, which is a histone acetyltransferase-coactivator complex involved in the global regulation of gene expression through acetylation and transcription functions.

Thus, our TF functional results are confirmed by the bibliography and, moreover, the literature shows a glucose-driven effect for some of the most relevant transcription factors (YLR403W and YGL209W). These two facts reinforce the reliability of our newly found functionalities. Considering that YMC includes glucose limited conditions, this behaviour perfectly fits with the kind of responses that would be expected by the significant variables regulating the YMC (they should be sensitive to nutrient changes in the environment). Among our results, the effect of Y0R363C is also linked to glucose starvation, which could be further analysed in future research. Furthermore, two of the most relevant transcription factors are also involved in the direct regulation of the chromatin status, as previously indicated, YPL254W (relevant for the SAGA complex) and YLR403W (involved in histone exchanges), which emphasizes the relevance of the study of both TFs and histone modifications for a better understanding of the YMC.

When studying the significant co-regulations of histone modifications and TFs, H3K18ac, one of the most relevant histone modifications regulating the gene expression in the YMC, was found to be significantly associated to YPL254W, a TF which is involved, as previously indicated, in the global regulation of the gene expression through the acetylation functionalities given by the SAGA complex. Thus, this does not seem to be an arbitrary relationship, but show the crucial role of the acetylation levels to regulate gene expression in the YMC through a coordinated action of transcription factors and histone modification levels.

Shortly, when comparing our results with the ones from (Kuang, et al., 2014), the project from which the data was obtained, we think that our integration strategy has helped us to improve or tune the results they got. However, it is true that not all the results are comparable, either because the goals of both studies were somehow different or because the data and experiments performed were not exactly the same. They were not able to assess clearly the relevance for the different studied histone modifications in the gene expression regulation of the YMC as done in this project, but we also must take into account that they did not included in their study one of the most important histone modifications, H3K18ac. Furthermore, in Kuang it is hypothesized that H3K14ac and H4K5ac have a role in pre-setting the promoters, perhaps by chromatin remodelling, despite there is a number of genes that do not show evidence of this preactivation phase from the prespective of chromatin states. Regarding our results H4K5ac does not present a notoriously relevant role in the regulation of the YMC, and it would not support the recently exposed hypothesis of Kuang. In addition, we presented an in-depth functional characterization of the functionalities regulated by each histone modification, which was lacking in previous studies. Moreover, the study of the transcription factors presented here is completely new. This offers a new perspective of the YMC, since nothing similar was done in the previous studies, and points towards regulators that can be further analysed in future studies. Hence, the methodology applied in this project has been able to improve and add new information to the results given in (Kuang, et al., 2014).

All in all, our results show the relevance of both transcription factors and histone modifications in the regulation of YMC gene expression. The literature not only supports our findings but also highlights the influence of different metabolic compounds on the cell chromatin status, which connects the metabolome with the levels of different histone modifications and with gene expression changes. Thus, the future generation of metabolomics data for this study should be considered, since it would contribute to gain more insights on the YMC regulation. Moreover, since the relevant studied TFs have been found to be important in the regulation of the YMC, the 7 remaining relevant TFs not analysed, due to time constraints, should be further analysed. Furthermore, studies of chromatin occupancy would help to clarify the role of these regulatory proteins in gene expression cooperating with histone modifications.

5. Conclusion

The contributions of this work are framed in two different areas, the biological and the methodological fields.

Regarding the methodological aspects, a novel application of different statistical methodologies for the integration of chromatin status data (ChIP-Seq) and gene expression (RNA-Seq) has been described, based on the usage of N-PLS and MORE techniques. The strategy followed here has demonstrated its utility since we have been able to assess the effect of different studied regulators (transcription factors and histone modifications) on gene expression changes across the YMC, and evidences have been found in the literature that support the results obtained. Hence, an effective and profitable bioinformatics strategy for the omics integration has been detailed in the field of the large scale omics data, which could be included in big data analysis category, due to the large amount of studied variables.

The work plan established at the beginning of the project was fulfilled satisfactorily. The modifications introduced along the thesis were done as a consequence of the natural evolution of the analysis, where new ideas can appear which lead to the addition of new steps that contribute to the improvement of the results. In our case, the main additions included the analysis of transcription factors and the creation of a regulatory network. However, a great effort, more than expected, was devoted to pre-process all the ChIP-Seq raw data, since the available data had been generated through different technologies and with different qualities, so many issues came up during the pre-processing step. The type of problems raised during data preparation are commonly overlooked, because they indeed consume a large amount of time, but can notoriously influence the results of the study. Given that these steps are so remarkably important, preventive measures should be taken, not only while preparing the data, but also during the design of the experiment, since in many occasions it is not possible to solve the data problems associated to a wrong experimental design, in which different effects are confounded.

Focusing on the biological results of the project, it has been found how DNA histones acetylation changes, and different transcription factors, present a significant role in the regulation of the YMC gene expression changes. H3K18ac and H3K9ac result as the most relevant of the studied histone modifications. YLR403W, YPL254W, YOR363C, YGL209W and YDR451C appear as the most relevant transcription factors. Moreover, along the discussion, a clear interaction between the metabolome, the chromatin status and the gene expression changes has been described, which has indicated how helpful it would be, in future steps, the addition of metabolomics data to improve the comprehension of the regulation in the YMC. In conclusion, it has been proven how the careful data pre-processing and the novel integrative approaches presented in this project have improved, or at least corroborated, previously reported results that were generated by independently analysing each omic data type.

6. Glossary

ANOVA	Analysis of variance
bp	Base pairs
DE	Differentially expressed
DEGs	Differentially Expressed Genes
FE	Functional Enrichment
GLM	Generalized Linear Model
GO	Gene Ontology
Nr.	Number
OX	Oxidative
PC	Principal Component
PCA	Principal Component Analysis
RB	Reductive Building
RC	Reductive Charging
SCA	Simultaneous Component Analysis
TF	Transcription Factor
TSS	Transcription Start Site
TTS	Transcription Termination Site

7. Bibliography

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

Baumgartner, U. et al., 1999. Functional analysis of the Zn(2)Cys(6) transcription factors Oaf1p and Pip2p. Different roles in fatty acid induction of beta-oxidation in Saccharomyces cerevisiae.. *The Journal of biological chemistry,* 6 8, 274(32), pp. 22208-16.

Benjamini, Y. & Hochberg, Y., 1995. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* s.l.:WileyRoyal Statistical Society.

Berger, S. L., 2007. The complex language of chromatin regulation during transcription. *Nature,* 24 5, 447(7143), pp. 407-412.

Bolger, A. M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 1 8, 30(15), pp. 2114-2120.

Bro, R., 1996. Multiway calibration. Multilinear PLS. *Journal of Chemometrics,* 1 1, 10(1), pp. 47-61.

Cai, L., Sutter, B. M., Li, B. & Tu, B. P., 2011. Acetyl-CoA induces cell growth and proliferation by promoting the acetylation of histones at growth genes.. *Molecular cell,* 20 5, 42(4), pp. 426-37.

Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J., 2016. Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics,* 9, 17(5), pp. 891-901.

Conesa, A., Nueda, M. J., Ferrer, A. & Talon, M., 2006. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics,* 1 5, 22(9), pp. 1096-1102.

Conesa, A. et al., 2010. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics and Intelligent Laboratory Systems,* 15 11, 104(1), pp. 101-111.

Coppi, R., Bolasco, S. & Consiglio nazionale delle ricerche (Italy), 1989. *Multiway data analysis.* s.l.:North-Holland.

Dopazo, J., 2006. Functional Interpretation of Microarray Experiments. *OMICS: A Journal of Integrative Biology,* 9, 10(3), pp. 398-410.

Draper, N. R. & Smith, H., 1998. *Applied regression analysis.* s.l.:Wiley.

Edgar, R., Domrachev, M. & Lash, A. E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.. *Nucleic acids research,* 1 1, 30(1), pp. 207-10.

Faraway, J. J., 2005. *Linear models with R.* s.l.:Chapman & Hall/CRC.

Fernández-Cid, A., Riera, A., Herrero, P. & Moreno, F., 2012. Glucose levels regulate the nucleo-mitochondrial distribution of Mig2.. *Mitochondrion,* 5, 12(3), pp. 370-80.

Gene Ontology Consortium, M. A. et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research,* 1 1, 32(90001), pp. 258D-261.

Genomics of Gene Expression group, n.d. *ConesaLab / more — Bitbucket.* [Online]
Available at: https://bitbucket.org/ConesaLab/more/src/master/

Hoerl, A. E. & Kennard, R. W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Source: Technometrics,* 12(1), pp. 55-67.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes.. *Nucleic acids research,* 1 1, 28(1), pp. 27-30.

Karunanithi, S. & Cullen, P. J., 2012. The filamentous growth MAPK Pathway Responds to Glucose Starvation Through the Mig1/2 transcriptional repressors in Saccharomyces cerevisiae.. *Genetics,* 1 11, 192(3), pp. 869-87.

Kuang, Z. et al., 2014. High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast.. *Nature structural & molecular biology,* 10, 21(10), pp. 854-63.

Kunoh, T., Kaneko, Y. & Harashima, S., 2000. YHP1 encodes a new homeoprotein that binds to the IME1 promoter in Saccharomyces cerevisiae.. *Yeast (Chichester, England),* 30 3, 16(5), pp. 439-49.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.. *Genome biology,* 10(3), p. R25.

Liang, K. & Keleş, S., 2012. Normalization of ChIP-seq data with control. *BMC Bioinformatics,* 10 8, 13(1), p. 199.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools.. *Bioinformatics (Oxford, England),* 15 8, 25(16), pp. 2078-9.

Marion, R. M. et al., 2004. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression.. *Proceedings of the National Academy of Sciences of the United States of America,* 5 10, 101(40), pp. 14315-22.

McCormack, M. E., Lopez, J. A., Crocker, T. H. & Mukhtar, M. S., 2016. Making the right connections: Network biology and plant immune system dynamics. *Current Plant Biology,* 1 4, Volume 5, pp. 2-12.

Mentch, S. et al., 2015. Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. *Cell Metabolism,* 11, 22(5), pp. 861-873.

Nelder, J. A. & Wedderburn, R. W. M., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) J. R. Statist. Soc. A,* 135(3), pp. 370-384.

Nikolsky, Y. et al., 2009. Functional Analysis of OMICs Data and Small Molecule Compounds in an Integrated “Knowledge-Based” Platform. In: s.l.:s.n., pp. 177-196.

Nueda, M. j., Ferrer, A. & Conesa, A., 2012. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics,* 1 7, 13(3), pp. 553-566.

Nueda, M. J., Tarazona, S. & Conesa, A., 2014. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics,* 15 9, 30(18), pp. 2598-2602.

Pramila, T. et al., 2002. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle.. *Genes & development,* 1 12, 16(23), pp. 3034-45.

Python Software Foundation, n.d. *Python is a programming language that lets you work quickly and integrate systems more effectively..* [Online]
Available at: https://www.python.org/

Quinlan, A. R. & Hall, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 15 3, 26(6), pp. 841-842.

R Development Core Team, 2011. *R: a language and environment for statistical computing.* [Online]
Available at: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

Ringnér, M., 2008. What is principal component analysis?. *NATURE BIOTECHNOLOGY,* 26(3).

Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A., 2017. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology,* 3 11, 13(11), p. e1005752.

Rottensteiner, H. et al., 1996. Pip2p: a transcriptional regulator of peroxisome proliferation in the yeast Saccharomyces cerevisiae.. *The EMBO journal,* 17 6, 15(12), pp. 2924-34.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 1 11, Volume 20, pp. 53-65.

Shannon, P. et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks.. *Genome research,* 1 11, 13(11), pp. 2498-504.

Smilde, A. K., Bro, R. & Geladi, P., 2004. *Multi-way analysis with applications in the chemical sciences.* s.l.:J. Wiley.

Tarazona, S. et al., 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research,* 16 7, 43(21), p. gkv711.

Teixeira, M. C. et al., 2018. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Research,* 4 1, 46(D1), pp. D348-D353.

Tibshirani, R., 1996. *Regression Shrinkage and Selection via the Lasso.* s.l.:WileyRoyal Statistical Society.

Tu, B. P., Kudlicki, A., Rowicka, M. & McKnight, S. L., 2005. Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science,* 18 11, 310(5751), pp. 1152-1158.

Wellen, K. E. et al., 2009. ATP-citrate lyase links cellular metabolism to histone acetylation.. *Science (New York, N.Y.),* 22 5, 324(5930), pp. 1076-80.

Wold, S., Sjöström, M. & Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 28 10, 58(2), pp. 109-130.

Xu, Z. & Norris, D., 1998. The SFP1 gene product of Saccharomyces cerevisiae regulates G2/M transitions during the mitotic cell cycle and DNA-damage response.. *Genetics,* 12, 150(4), pp. 1419-28.

Yu, D., Kim, M., Xiao, G. & Hwang, T. H., 2013. Review of biological network data and its applications.. *Genomics & informatics,* 12, 11(4), pp. 200-10.

Zerbino, D. R. et al., 2018. Ensembl 2018. *Nucleic Acids Research,* 4 1, 46(D1), pp. D754-D761.

Zhao, S. et al., 2010. Regulation of cellular metabolism by protein lysine acetylation.. *Science (New York, N.Y.),* 19 2, 327(5968), pp. 1000-4.

Zhou, X. & Su, Z., 2007. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species.. *BMC genomics,* 24 7, Volume 8, p. 246.

Zou, H. & Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B,* 67(2), pp. 301-320.