UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Agricultural Engineering and Environment

BENCHMARKING OF DIFFERENT OMIC TECHNOLOGIES FOR METABOLIC MODELING USING CONSTRAINT- BASED AND FUNCTIONAL ENRICHMENT-BASED METHODS

End of Degree Project

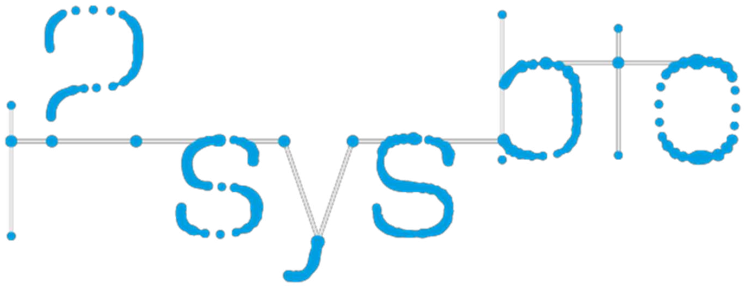
Bachelor's Degree in Biotechnology

AUTHOR: Araiz Sancho, Cristina

Tutor: López Gresa, María Pilar

External cotutor: CONESA CEGARRA, ANA

ACADEMIC YEAR: 2022/2023



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Agricultural Engineering and Environment

Benchmarking of different Omic Technologies for Metabolic Modelling using Constraint-based and Enrichment-based methods

BACHELOR'S DEGREE IN BIOTECHNOLOGY

Final Degree Project

Author: Cristina Araiz Sancho

Tutors: Ana Conesa Cegarra

María Pilar López Gresa

Advisor: Alejandro Paniagua de Pedro

Academic year: 2022 - 2023

Valencia, June 10, 2023

Title

Benchmarking of different Omic Technologies for Metabolic Modelling using Constraint-based and Functional Enrichment-based Methods

Título

Estudio Comparativo de diferentes Tecnologías Ómicas para el Modelado de Redes Metabólicas a partir de Métodos de Restricción y Enriquecimiento Funcional

Títol

Estudi Comparatiu de diferents Tecnologies Ómiques per al Modelatge de Xarxes Metabòliques a partir de Mètodes de Restricció i Enriquiment Funcional

Abstract

Multi-omic approaches, which combine information from various biological data layers, have become increasingly popular in recent years because of their ability to provide a thorough understanding of biological systems.

This study aims to conduct a comprehensive evaluation of MAMBA (Metabolic Adjustment via Multiomic Blocks Aggregation), a Constraint-Based Modeling approach recently developed in the Conesa Lab that extends Flux Balance Analysis by incorporating biological data from different molecular layers. *Saccharomyces cerevisiae*, specifically the Yeast Metabolic Cycle, provides an interesting biological scenario for this purpose due to its well-annotated genome and abundance of experimental data.

The corresponding multi-omic datasets were retrieved from public repositories, and include transcriptomic, epigenomic and metabolomic information for 16 time points distributed throughout the 3 phases of the cycle. The data were obtained using RNA-seq, ChIP-seq, and LC-MS techniques, respectively.

Differential expression analysis was independently applied to each omic. Further, functional analysis by means of conventional enrichment-based methods (ORA, GSEA) and using the novel constraint-based approaches (MAMBA) allowed to evaluate the performance of both types of approaches for the different omics. ORA and GSEA identified metabolic pathways that contained differentially expressed genes across the three phases, aligning with previous knowledge on the Yeast Metabolic Cycle. The reaction activity prediction feature implemented in MAMBA also identified differentially active pathways, albeit with only partial overlap between the pathways identified by the enrichment methods. Enrichment analysis exposed regulation along the yeast metabolic cycle for critical biological functions, such as DNA replication, proteasome degradation, and mismatch repair. These differences were not observed in MAMBA's results as it is confined to the metabolic reactions within the model. However, MAMBA's results were more proficient in capturing the downstream effects of these fundamental processes, represented by specific reactions and metabolic pathways.

In order to evaluate the performance of both methods in terms of identifying and capturing the biological differences between conditions, a Gene Set Variation Analysis (GSVA) was applied only to those pathways identified as relevan. The resulting GSVA scores were then utilized to cluster the samples in an unsupervised manner. In the case of RNA-seq, pathways identified by both MAMBA and ORA accurately captured the underlying biological differences between conditions for both MAMBA and ORA. Notably, MAMBA exhibited superior performance when employing epigenetic information.

In conclusion, MAMBA shows great potential as an effective approach for integrating transcriptomic and epigenetic data to unravel the complexities of biological processes such as the yeast metabolic cycle. However, further research is needed to enhance the functional characterization of the metabolic model and optimize the incorporation of diverse omics datasets within the MAMBA framework. These efforts will contribute to improving the accuracy and reliability of MAMBA's predictions, thereby advancing our understanding of complex biological systems.

This project is related to the Sustainable Development Goals, specifically contributing to Health and Well-being (SDG 3), Quality Education (SDG 4), Gender Equality (SDG 5), Climate Action (SDG 11) and Responsible production and consumption (SDG 12).

Keywords Multi-omic; Metabolic Modelling; Flux Balance Analysis; *Saccharomyces cerevisiae*; Yeast Metabolic Cycle; RNA-seq; Chip-seq; Metabolomics

Resumen

Los enfoques multi-ómicos, que combinan información de varias capas de datos biológicos, se han vuelto cada vez más populares en los últimos años debido a su capacidad para proporcionar una comprensión profunda de los sistemas biológicos.

Este estudio tiene como objetivo llevar a cabo una evaluación exhaustiva de MAMBA, una herramienta de integración multi-ómica que extiende el Análisis de Balance de Flujos mediante al incorporación conjunta de datos biológicos de diferentes capas moleculares. *Saccharomyces cerevisiae*, en concreto el Ciclo Metabólico de la Levadura, ofrece un escenario biológico interesante para este propósito debido a la extensa caracterización de su genoma y a la abundancia de datos experimentales. Los datos multi-ómicos correspondientes se extrajeron de repositorios públicos e incluyen información transcriptómica, epigenómica y metabolómica para 16 puntos de tiempo distribuidos a lo largo de las 3 fases del ciclo.

El análisis de expresión diferencial se aplicó de manera independiente a cada ómica. Posteriormente, el análisis funcional mediante métodos convencionales de enriquecimiento (Over Representation Analysis, Gene Set Enrichment Analysis) y mediante los nuevos métodos basados en restricciones (MAMBA) permitió realizar una evaluación comparativa del rendimiento de ambas metodologías.

ORA y GSEA identificaron vías metabólicas enriquecidas en genes diferencialmente expresados a lo largo de las tres fases. MAMBA también permitió identificar vías diferencialmente activas a partir de la predicción de flujos. El análisis de enriquecimiento permitió identificar los procesos biológicos clave en cada fase del ciclo, tales como la replicación del ADN o la síntesis de proteinas. Estas funciones globales no se observaron en los resultados de MAMBA, limitados a las reacciones metabólicas explícitamente incluidas en el modelo. Sin embargo, MAMBA demostró capturar con éxito los efectos aguas abajo de estos procesos fundamentales, en forma de reacciones y vías metabólicas específicas.

Con el objetivo de evaluar si ambos métodos eran capaces de capturar las diferencias biológicas entre las fases, se realizó un Análisis de Variación de Conjunto de Genes incluyendo únicamente las vías metabólicas identificadas como relevantes en los análisis previos. Las puntuaciones resultantes se utilizaron para agrupar las muestras de manera no supervisada. Pese a que en el análisis de los datos transcriptómicos no se observó una diferencia significativa entre el ORA y MAMBA, siendo ambos capaces de identificar y resumir las diferencias biológicas fundamentales entre ambos métodos, MAMBA demostró un desempeño claramente superior en el análisis basado en datos datos epigenéticos.

En conclusión, MAMBA emerge como una herramienta con gran potencial en la integración de datos epigenéticos y transcriptómicos para estudiar procesos biológicos complejos, tales como el Ciclo Metabólico de la Levadura. Sin embargo, sería conveniente optimizar la caracterización del modelo metabólico y desarrollar adaptaciones de MAMBA más específicas que tengan en cuenta la naturaleza de las distintas capas moleculares.

Este proyecto se enmarca dentro de los Objetivos de Desarrollo sostenible, en concreto contribuyendo a la Salud y Bienestar (ODS 3), Educación de calidad (ODS 4), Igualdad de Género (ODS 5), Acción por el clima (ODS 11) y Producción y consumo responsables (ODS 12).

Palabras clave Multi-ómica; Modelado Metabólico; Análisis de Balance de Flujos; *Saccharomyces cerevisiae*; Ciclo Metabólico de la Levadura; RNA-seq; ChIP-seq; Metabolómica

Acknowledgments

First and foremost, I would like to express my gratitude to the entire team at the Genomics of Gene Expression Lab:

To Ana Conesa, for granting me the opportunity to carry out my Bachelor's Thesis in this laboratory.

In particular, I would like to acknowledge the exceptional support of Alejandro, who has been by my side from the first day until the last. Without his infinite patience, valuable ideas, dedication, and commitment, the outcome of this project would have been completely different.

Furthermore, I would like to thank all my colleagues at the laboratory, who made working there not only academically enriching but also enjoyable and pleasant from the very beginning.

I am also very grateful to Mapi for her role as both teacher, where she was able to inspire us with enthusiasm for her subject, and supervisor, where she has diligently supervised the project with great interest from the beginning.

Secondly, I would like to express my gratitude to my family for standing by me during the challenging moments. Especially to my mother, for instilling in me a passion for science and biology, and for teaching me from a young age the importance of being curious and critical to understand the world around us.

Lastly, I want to thank my friends from university (María, Alejandro, Miguel, Juanto, Carmen, and Adri). I have been truly happy during these four years, and this is largely thanks to all of you. Although we may be concluding this chapter together, I leave with the hope and confidence that you will continue to be by my side in the future.

Contents

List of Figures

List of Tables

Acronyms

YMC	Yeast Metabolic Cycle
OX	Oxidative
RB	Reductive Builiding
RC	Reductive Charging
GSEA	Gene Set Enrichment Analysis
ORA	Over Representation Analysis
MAMBA	Metabolic Adjustment via Multi-Blocks Aggregation Method
CV	Coefficient of Variation
PCA	Principal Component Analysis
GSVA	Gene Set Variation Analysis
H3K9ac	Histone 3 Lysine 9 acetylation
H3K18ac	Histone 3 Lysine 18 acetylation
ChIP-seq	Chromatin Inmunoprecipitation Sequencing
RNA-seq	Ribonucleic Acid Sequencing
FE Analysis	Functional Enrichment Analysis
DE Analysis	Differential Expression Analysis
DE genes	Differentially Expressed genes
DAR	Differentially Active Reaction

1 Introduction

1.1 Multi-omic

Over the past few decades, there has been a remarkable emergence of high-throughput technologies that offer the opportunity of massively analysing diverse data derived from the different biochemical components of biological systems. The study of these omics data across different molecular layers of cellular processes help to understand the complex connections underlying the genetic information and the resulting observable traits (Marín de Mas, 2018). Some of the most widely used omics in systems biology are epigenomics, transcriptomics and metabolomics (Fig. 1)(Ugidos Guerrero, 2023).

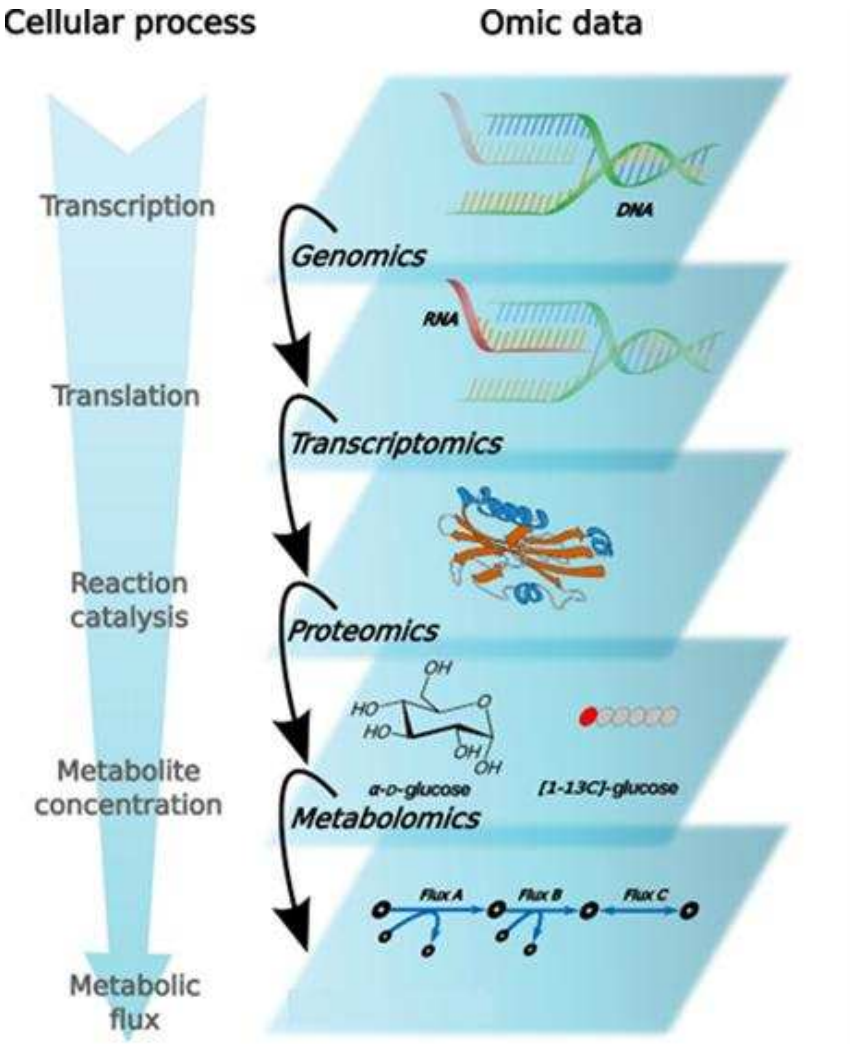


Figure 1: Omic technologies. The most relevant omics data in system biology represented as the different layers of biological information (Figure adapted from Marín de Mas, 2018).

Transcriptomics encompasses the comprehensive investigation of all transcripts present within a cell, involving their identification and quantification under specific biological states or condi-

tions (Wang *et al.*, 2009). To analyze the transcriptome, various technologies can be employed, including hybridization-based and sequencing-based approaches, with RNA-sequencing (RNA-seq) currently being the most prevalent method. In essence, RNA-seq involves the conversion of an RNA population into a library of cDNA fragments that have adaptors attached to their ends. Subsequently, high-throughput sequencing technologies are employed to generate short reads from either one or both ends of these cDNA fragments.

Epigenetics is a field of study that investigates alterations to the genome that do not involve changes in the underlying DNA sequence. These modifications can be broadly classified into two categories: nucleotide modifications and histone modifications. Both types of modifications play a pivotal role in the regulation of gene expression by influencing the packaging of DNA and modifying the surface of nucleosomes (O'Geen *et al.*, 2011).

However, due to the diverse range of alteration types (such as methylation and acetylation), the multitude of potential histone modification sites, and the possibility of simultaneous modifications, comprehending their precise impact on gene expression regulation is inherently challenging.

Chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) is a widely employed technique for investigating histone marks on a genome-wide scale. Briefly, ChIP-seq involves the initial crosslinking of DNA-protein complexes, followed by chromatin fragmentation and subsequent enrichment of the fragments using antibodies specific to the histone modifications of interest. The crosslinks are then reversed, and the DNA fragments are sequenced (O'Geen *et al.*, 2011).

Metabolomics is an emerging scientific field that aims to comprehensively measure all metabolites and low-molecular-weight molecules present in a biological sample. Unlike genomic and proteomic, metabolomics aims to study molecules that have very different physical properties, which represents a significant analytical challenge (Clish, 2015).

Analytical platforms commonly utilized in metabolomics include Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR). While NMR may have lower sensitivity compared to MS, it offers valuable advantages such as high reproducibility, non-selective detection and non-invasive nature (Gowda & Raftery, 2021).

By employing multi-omic approaches that encompass the simultaneous analysis of various omic data types, a deeper understanding of the cell as a complex system can be achieved.

1.2 Functional Enrichment Methods

The advent of high-throughput omic technologies has facilitated the measurement of gene expression levels in a biological sample of interest. However, comprehending the underlying molecular phenomena remains a significant challenge. In many studies focusing on the identification of significant differences between two groups or conditions of interest, a preliminary step involves conducting a Differential Expression (DE) Analysis. DE Analysis allows to identify which genes expressed at distinct levels between the groups, providing insight into the cellular processes affected by the different conditions. DE analysis returns a list of genes along with their corresponding Fold Change (FC) or magnitude of change and statistical measures like p-values, but these results alone may be challenging to interpret in terms of biological significance.

Functional enrichment (FE) analysis provides a valuable solution to this issue. FE methods rely on previous biological knowledge that enables to include each gene within pre-define "gene sets", that is, collection of genes sharing a certain characteristic (biological process, location, disease or a given metabolic pathway).

One of the most basic forms of FE analysis is Over Representation Analysis (ORA). ORA is a statistical method that determines whether genes from pre-defined gene sets are present more than expected (overrepresented) in a collection of differentially expressed (DE) genes. The underlying principle is that a pathway exhibiting a significantly higher occurrence of DE genes

than expected is more likely to be biologically relevant to the given condition. Given a list of significantly differentially expressed genes (L), and a gene set (G) that has n genes in common with L, ORA considers G as being differentially enriched if the occurrence of n in G is higher than would be expected by chance (Maleki *et al.*, 2020). For each G, an enrichment p-value is calculated one-by-one in a linear mode (Huang *et al.*, 2009) using the Binomial distribution, Hypergeometric distribution, the Fisher exact test, or the Chi-square test.

However, the main limitation of ORA is its dependence on an arbitrarily defined threshold to select the list of DE genes. Consequently, the results are highly influenced by the criteria employed to classify genes as DE, including the choice of statistical tests and magnitude or significance thresholds.

In this context, Functional Class Scoring methods, such as Gene Set Enrichment Analysis (GSEA) (Fig. 2) provide an interesting alternative, as they eliminate this dependency on gene selection criteria by taking all gene expression values into consideration.

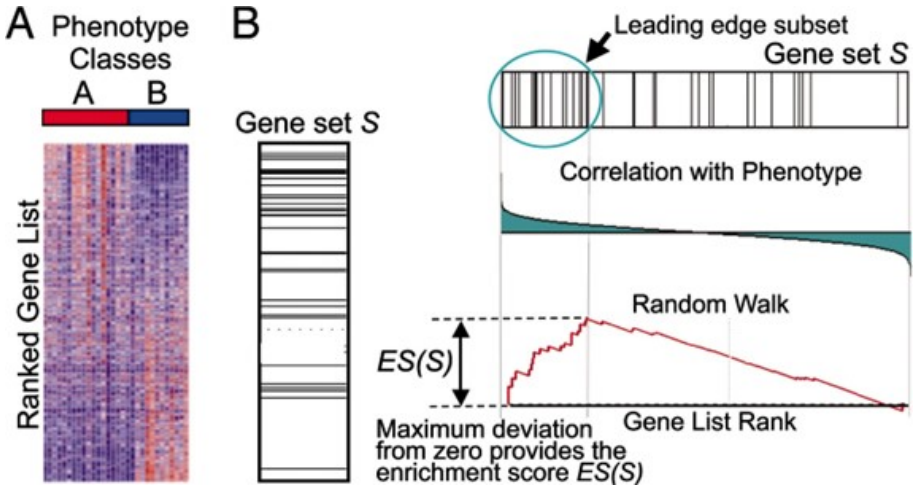


Figure 2: Overview of Gene Set Enrichment Analysis. (A) Genes are ranked according to a statistical metric accounting for differential expression between conditions. (B) Enrichment Score (ES) is calculated for each pathway based on the distribution of the corresponding genes throughout the ranked list (Adapted from Subramanian *et al.*, 2005).

GSEA, similar to ORA, evaluates gene expression data at the level of gene sets. It takes as input a ranked list of genes (L) based on a metric derived from the differential expression analysis. If a gene set (S) is unrelated to the phenotypic differences between the experimental conditions being compared, the genes in S are expected to be randomly distributed throughout L. Conversely, if S is associated with the phenotypic differences, the corresponding genes are anticipated to predominantly occur at the top or bottom of the list (Subramanian *et al.*, 2005).

To examine this distribution, GSEA computes an enrichment score (ES) for each gene set by walking down the list L, increasing a running-sum statistic when a gene belonging to S is encountered and decreasing it when the gene is not in S. For each ES, defined as the maximum deviation from zero encountered in the random walk (Fig. 2), its significance level is estimated (nominal p-value) and adjusted for multiple hypothesis testing (Subramanian *et al.*, 2005).

GSEA, despite proving a more unbiased approach for functional analysis, still relies on the pre-defined gene sets. Considering metabolic pathways as un-order and unstructured collection of genes discards a substantial amount of information about the biological processes described by these pathways, as they do not contemplate the dependencies and interaction between genes or other cellular components (Fig. 3).

Therefore, alternative methodologies that incorporate more comprehensive information about genes, for example in form of metabolic models, may prove valuable in attaining a deeper understanding of biological processes.

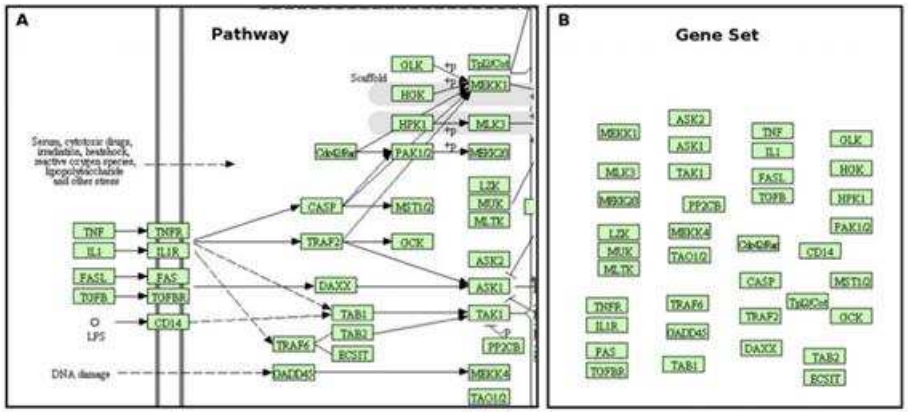


Figure 3: Gene set vs pathway analysis. Schematic representation of the information that is lost when considering metabolic pathways as gene sets (Adapted from ADVAITA, 2019).

1.3 Constraint-Based Modeling

MAMBA (Metabolic Adjustment via Multiomic Blocks Aggregation) is a Constraint-Based Modeling (CBM) approach recently developed in the Conesa Lab that extends Flux Balance Analysis (FBA) by incorporating biological data from different molecular layers, such as transcriptomics, epigenomics, or metabolomics.

1.3.1 Flux Balance Analysis

FBA is a mathematical approach widely used to study the flow of metabolites through a metabolic network. It is based on genome-scale metabolic reconstructions, which encompass all known reactions in an organism and the associated genes that code for the enzymes (Orth *et al.*, 2010). The metabolic network is represented mathematically as a stoichiometric matrix (Fig. 4), where each row corresponds to a unique metabolite, and each column represents a reaction. The stoichiometric coefficients in the matrix are negative for metabolites consumed, positive for metabolites produced, and zero if a metabolite does not participate in a reaction. The fluxes through the network reactions are represented by the vector x.

These stoichiometric relationships impose mass balance constraints on the system by ensuring that the total amount of any compound being produced is equal to the total amount being consumed, thereby maintaining the steady state. The system does not have a unique solution, and any flux distribution (x) that satisfies the mass balance equation lies in the null space of the stoichiometric matrix. In FBA, additional constraints in the form of upper and lower bounds can be added to each reaction, further limiting the allowable flux distributions.

To narrow down the solution space, FBA incorporates a biological objective that represents a relevant function to the problem being studied, such as biomass production. The objective is included as an additional reaction to the stochiometric matrix, where metabolites important for biomass production have negative coefficients, indicating their consumption during cell growth.

Thus, an objective function, in which all reactions have a zero coefficient except for the biomass production reaction, is stablished. The system of linear equations defined by the stoichiometric matrix and constraints is solved using linear programming techniques. The optimization problem aims to find the flux distribution that maximizes the objective function while satisfying the mass balance and flux constraints.

1.3.2 MAMBA

In MAMBA, the basic FBA is extended by incorporating biological data to determine an optimal metabolic flux solution that aligns with a specific phenotype of interest (Fig. 5). This integration

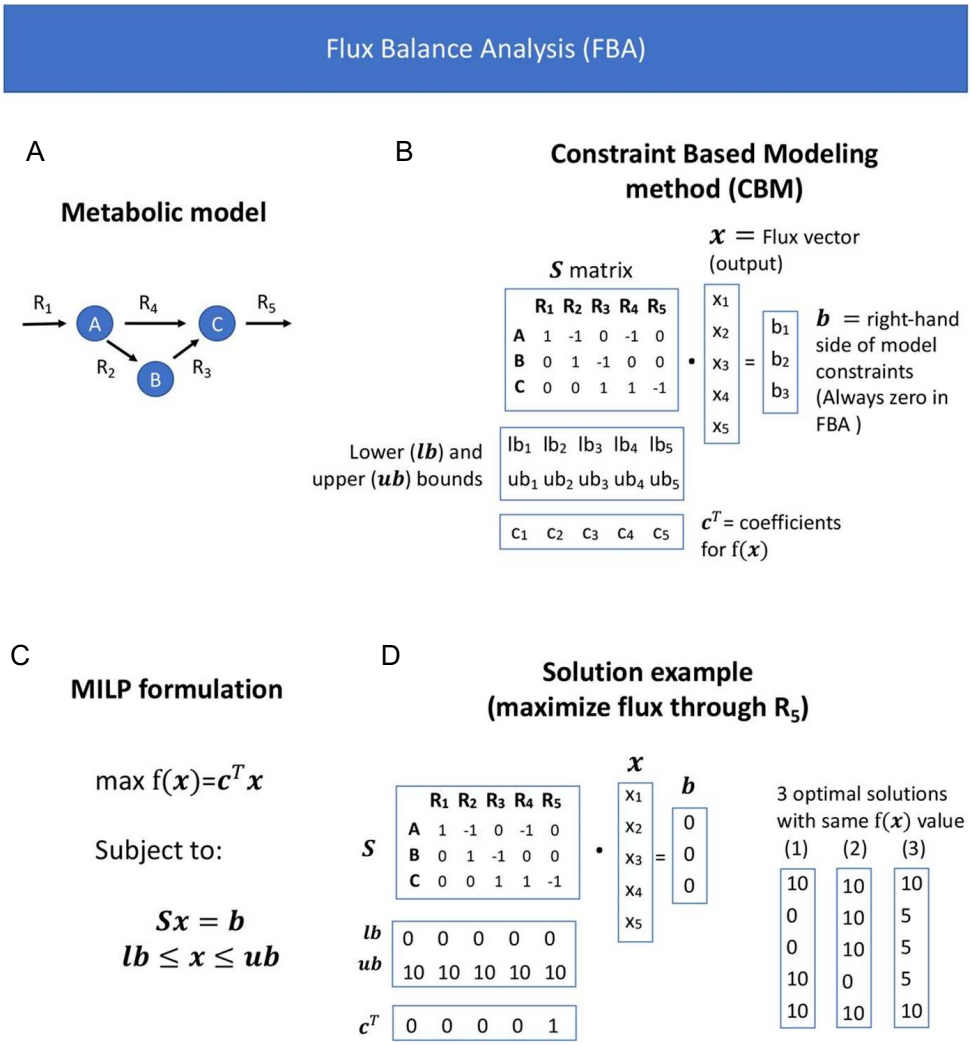


Figure 4: Overview of Flux Balance Analysis. (A) A simplified version of the metabolic network Representation of the metabolic network. (B) Mathematical representation of FBA elements, including Stochiometric matrix (S), Flux vector (x), steady state assumption (b), coefficients of the objective function for each reaction (C) and capacity constraints (lower and upper bounds). c) FBA formulated as a linear programming problem. (D) Example of a FBA solution for the network (Adapted from Ugidos *et al.*, 2022).

of gene-centric information is achieved through the implementation of Gene-Protein-Reaction rules (GPRs), which describe the associations between genes, proteins, and the corresponding metabolic reaction.

Gene-centric data, such as transcriptomic (RNA-seq) or epigenetic (ChIP-seq) data, are utilized to define the activation state of genes. MAMBA overcomes the challenge of setting absolute threshold values for gene states by working with differential values between compared conditions rather than absolute values (Ugidos *et al.*, 2022). An effect-size measure, such as FC, along with associated p-values resulting from a DE analysis, allow to determine whether a gene is active or inactive in a particular condition based on significant over- or underexpression with respect to other condition(s).

Based on the relative omics data, genes are categorized into three groups: UP (genes with significantly increased activity), DOWN (genes with decreased activity), and CONSTANT (genes with non-significant changes in activity) (Ugidos *et al.*, 2022). The incorporation of these gene states into the model is achieved through capacity constraints (upper and lower bounds) of the corresponding reactions. In a simplified scenario where the gene-reaction association is one-to-one, the following rules apply: for reactions associated with DOWN genes, the upper bound (Ub) and lower bound (Lb) are set to zero, effectively forcing the reaction to be inactive. For

UP genes, Ub and Lb are set to non-zero values, ensuring the reaction remains active. For CONSTANT genes, the reaction state is left undetermined (Lb zero and Ub non-zero), and it will adopt the state that best fits the remaining constraints in the model.

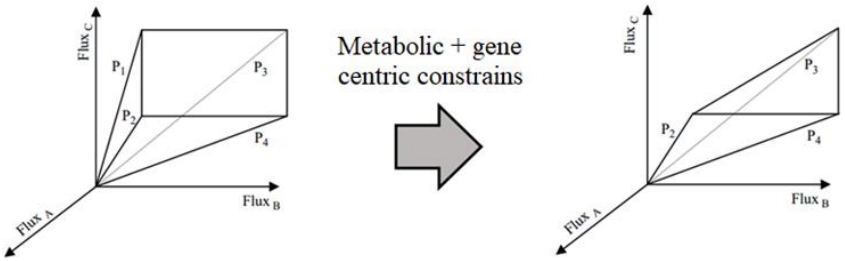


Figure 5: Constraints reduce solutions space. Omics data are incorporated as constrains to reduce the steady-state solution space of the metabolic network (Adapted from Covert & Palsson, 2003).

The algorithm in MAMBA searches for a sequence of binary expression states (active/inactive) that optimally fits the differential expression data for each compared condition. By iteratively adjusting the gene states and solving the resulting FBA problem, MAMBA determines the metabolic flux distribution that aligns with the observed gene expression changes, allowing for a more accurate representation of metabolic adjustments in response to different conditions or perturbations (Fig. 5).

In addition to gene-centric data, MAMBA integrates semi-quantitative metabolomics information to enhance the accuracy of metabolic predictions. This information is incorporated as differential values between conditions and is used by MAMBA to optimize the flux distribution in order by minimizing the discrepancy between the predicted and experimentally measured metabolite ratios.

1.4 Yeast Metabolic Cycle

Saccharomyces cerevisiae was chosen as the biological scenario for evaluating conventional enrichment-based and model-based functional analysis approaches. *Saccharomyces cerevisiae*, or budding yeast, is a well-established model organism for this type of study due to its extensively annotated genome, well-characterized metabolic network, and the abundance of experimental data available in public repositories.

Budding yeast cells exhibit oscillatory dynamics in various cellular pathways, including those involved in cell cycle, glucose metabolism, and respiration. This metabolically driven rhythm is known as the Yeast Metabolic Cycle (YMC) (Rao & Pellegrini, 2011). Studies that link post-translational modifications of histones to cycling transcripts in the YMC have revealed that chromatin is globally altered as cells progress through the cycle, providing a valuable opportunity to investigate cell metabolism from a multi-omic perspective (Cai *et al.*, 2011).

Monitoring the YMC is achieved by measuring the dissolved oxygen levels in the media, which reflect the oxygen consumed by the cells. This allows for the integration of data from different studies or molecular data layers. The periodic oscillations in oxygen levels suggest that cells transition between two metabolic states: a high oxygen consumption phase and a low oxygen consumption phase (Mellor, 2016).

Previous studies on the functional profiling of the YMC reported that gene expression cycles in three distinct phases: the Oxidative phase (OX), Reductive Building phase (RB), and Reductive Charging phase (RC) (Fig. 6).

The OX phase is characterized by increased oxygen consumption and encompasses energetically demanding processes such as protein synthesis. Genes overexpressed in this phase encode

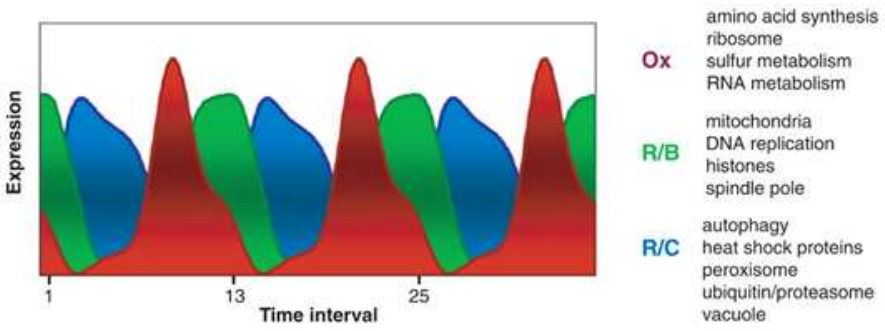


Figure 6: Gene expression oscillations along the YMC. Gene expression cycles in 3 distinct phases (OX, RB and RC), each of them with a characteristic functional profile (Adapted from Casaní Galdón, 2021).

for proteins involved in transcriptional machinery, ribosome biogenesis, and amino acid biosynthesis. The RB phase follows, with ongoing high oxygen consumption. Gene profiles during this phase exhibit a lower dynamic range and are mainly associated with DNA replication. Finally, the RC phase begins as oxygen levels start to increase and is characterized by non-respiratory metabolism and protein degradation (Casaní Galdón, 2021).

This data set, generated by Kuang *et al.* (2014) and Casaní Galdón (2021) and retrieved from public repositories, provided a comprehensive resource for investigating the YMC and its associated molecular dynamics.

Given the comprehensive nature of this multi-omic dataset, which offers insights into the cellular activity of the YMC in *Saccharomyces cerevisiae* at various molecular levels, it presents an ideal scenario for assessing the performance of MAMBA in metabolic modelling of a specific condition of interest. Furthermore, it allows for the exploration of MAMBA's potential in uncovering novel metabolic pathways involved in cellular processes that may not be readily captured by conventional enrichment-based methods.

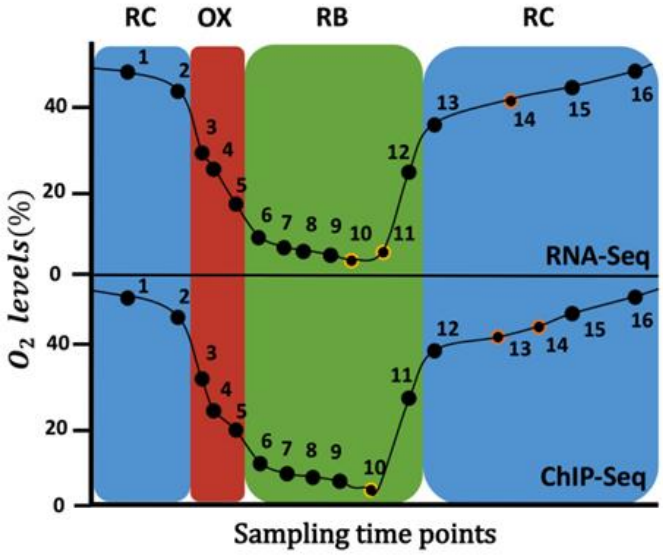


Figure 7: Time points sampled during the YMC. RNA-seq and of ChIP-seq samples used in this study are aligned and associated to the corresponding metabolic phase by measuring the dissolved oxygen levels (Adapted from Casaní Galdón, 2021).

2 Objectives

The primary objective of this bachelor's thesis is to conduct a comprehensive evaluation of MAMBA as a metabolic modelling tool, specifically focusing on its performance when applied to different types of omic data, including transcriptomic and epigenetic data. In addition, this thesis aims to characterize the advantages and limitations of MAMBA in comparison to more conventional enrichment-based methods commonly used in functional analysis. To achieve these objectives, the thesis will pursue the following main goals:

1. Pre-processing and Differential Expression analysis on a multi-omic dataset that encompasses transcriptomic, epigenomic, and metabolomic information.
2. Conduct functional analysis using functional enrichment methods to identify enriched biological functions associated with the differentially expressed genes. Compare the overregulated pathways inferred from the transcriptomic and epigenetic information to investigate potential differences.
3. Use MAMBA for metabolic modelling and perform functional analysis based on the flux predictions. Compare the functional profiles obtained from the different molecular layers to identify potential regulatory relationships.
4. Evaluate and compare the performance of both MAMBA and conventional enrichment-based methods in capturing the underlying biological differences between the studied conditions.

3 Materials and methods

3.1 Omic data

3.1.1 RNA-seq

RNA-seq data availability

Short-read RNA-seq data from *Saccharomyces cerevisiae* used in the expression profiling of the Yeast Metabolic Cycle (YMC), generated by Kuang *et al.* (2013), was downloaded from SRA accession SRP032970 using fasterq-dump. This included 16 time points over one metabolic cycle (GEO_accession: GSM1263339 to GSM1263354).

RNA-seq data pre-processing

RNA-seq data from the 16 time samples were pre-processed simultaneously. Initial quality control from the reads was done using fastqc (LaMar, 2015) with default parameters. Reads were trimmed to remove N content from the ends using cutadapt (Martin, 2011) with option *–trim-n*. Final quality check was performed with fastp (Chen *et al.*, 2018). The reference genome of *Saccharomyces cerevisiae* (Strain: S288C) and structural annotation was downloaded from RefSeq database (GCF_000146045.2). STAR v2.7.10 (Dobin *et al.*, 2013) was used for indexing the genome and aligning the processed reads, with option *–quantMode GeneCounts* to obtain raw counts for each gene.

3.1.2 ChIP-seq

Histone modification data generated by Kuang *et al.* (2013) and processed by Casaní (2021) were used to study epigenetic regulation. Only two out of the eight histone modifications in Casaní's study were considered (H3K9ac and H3K18ac), as they proved to have the greatest correlation with transcriptomic data (Casaní Galdón, 2021). H3 that was used as a control. Counts for each gene were obtained by calculating the average read coverage in a pre-defined region (300 bp upstream from the Transcription Start Site). TMM (Trimmed Mean of M values) normalization (Robinson *et al.*, 2010) using NOISeq R package was applied to make all samples have the same dynamic range (Ugidos Guerrero, 2023). Then, the average H3 count value of each gene for each phase was calculated. H3K9ac and H3K18ac were normalized with H3 control by dividing each gene count by the H3 mean for that gene in the corresponding phase.

3.1.3 Metabolomics

The normalized metabolomic values used in this study were obtained from Casaní (2021). Metabolic measurements were performed on a Liquid Chromatography - Mass Spectrometry (LC-MS) platform. After correction for protein abundance and median normalization, the final data set consisted of 50 metabolites in 21 time-points distributed along the cycle.

3.2 Study of variability

Variability of gene expression inside and between phases was evaluated by means of several complementary statistical methods, including Coefficient of Variation (CV), Principal Component

Analysis (PCA) and unsupervised clustering. All analyses were performed in RStudio.

3.2.1 Coefficient of Variation

CV testing was applied only to the RNA-seq dataset and the results were assumed to be representative of the rest of the omics. CV is defined as the ratio of the standard deviation to the mean. For gene expression values, the CV was calculated for each gene in each of the 3 phases (Eq. 1).

$$\mathrm{CV} = \frac{\mathrm{sd}}{\hat{x}} \tag{1}$$

In order to infer the significance of these values, 50 sample groupings were randomly generated maintaining the original size of the phases: 3 samples were randomly assigned to OX, 6 samples to RB and 6 to RC. The mean of CV (all genes) of each random grouping phase was compared with the CV mean of the original grouping. Finally, the average of the 3 phases in each random grouping was compared to the original biological-based grouping.

3.2.2 Clustering:

PCA was performed for all 4 omics data sets independently (RNA-seq, ChIP-seq H3K9, ChIP-seq H3K18, metabolomics). Hierarchical unsupervised clustering based on the gene-expression data was used as a complementary test.

3.3 Differential Expression Analysis

3.3.1 RNA-seq

In the case of RNA-seq, differential expression (DE) analysis was performed in RStudio using DESeq2 package, which tests for differential expression by use of negative binomial generalized linear models (Love *et al.*, 2014). Time point 12 was excluded from the analysis due to its lower resemblance to the other samples in RB, according to the results of the previous study of variability.

Two objects were created for this purpose: count matrix (raw counts of the 15 time samples, non-normalized) and sample information (YMC phase assigned to each sample in order to define the groups to be compared). The contrasts were done as pairwise comparisons considering the biologically meaningful transitions: OX vs RC (from RC to OX), RB vs OX (from OX to RB) and RC vs RB (from RB to RC).

DE Analysis output for each contrast (phase transition) consisted of six columns of information reported for each gene: mean of normalized counts for all samples, log2FoldChange (log2FC), standard error, Wald statistic, p-value and Benjamini-Hochberg adjusted p-value (p-adj), accounting for multiple testing. Those genes with a p-adj < 0.05 were considered significant.

3.3.2 ChIP-seq and metabolomics

In the context of ChIP-seq and metabolomic data analysis, the LIMMA package (Ritchie *et al.*, 2015) was employed to conduct DE analysis of gene counts and metabolite abundance in RStudio. The normalized counts of the two histone modifications (H3K9ac and H3K18ac) were analysed separately. Similar to the approach utilized in RNA-seq analysis, the samples were divided into three groups based on the YMC phases, and pairwise comparisons were conducted considering the direction of the cell cycle (i.e., OX vs. RC, RB vs. OX, and RC vs. RB).

Several statistics were computed for each gene and contrast, including log2FC, standard errors, t-statistics, p-value and Benjamini-Hochberg p-adj (accounting for multiple testing). Those genes with a p-adj < 0.05 were considered significant.

3.4 Functional analysis with enrichment – based methods

To gain a more comprehensive understanding of the differentially expressed (DE) genes, the outcomes of the DE analysis were examined using two distinct functional enrichment methodologies, namely Gene Set Enrichment Analysis (GSEA) and Over-Representation Analysis (ORA). The annotation of genes was carried out utilizing Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database.

3.4.1 Gene Set Enrichment Analysis

GSEA requires a ranked list of genes derived from differential expression analysis. The choice of ranking metric for GSEA significantly influences the outcomes of pathway enrichment analysis, as emphasized by Zyla *et al.* (2017). In this particular study, the genes were ranked based on the Wald Statistic, which accounts for both the significance and the direction/magnitude of the gene expression changes in the cycle's phase transitions.

For GSEA, the WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) platform was employed, with the following parameter settings: organism (*Saccharomyces cerevisiae*), ID type (ensemble_gene_id), functional database (KEGG pathway), minimum number of IDs (20), maximum number of IDs (500), False Discovery Rate (FDR) method (Benjamini-Hochberg), significance level (FDR < 0.05) and the number of permutations (1000).

3.4.2 Over Representation Analysis

ORA was independently conducted for each phase, exclusively considering genes deemed significant (p-adj < 0.05) and exhibiting upregulation in that phase with respect to the other two. This selection process took into account both the significance (p-adj) and the direction (log2FC) of the change, respectively.

ORA was carried out using the WebGestalt platform to analyse the three phases and the three omics independently. The analysis incorporated the following parameter settings: organism (*Saccharomyces cerevisiae*), ID type (ensemble_gene_id), functional database (KEGG pathway), minimum number of IDs (5), maximum number of IDs (2000), FDR Method (Benjamini-Hochberg), and significance level (FDR < 0.05). The entire list of genes encompassed in the sequencing experiment was utilized as the background for the analysis.

Those pathways with a FDR value lower than 0.05 were considered biologically relevant.

3.5 MAMBA

3.5.1 Flux prediction

To streamline the computational analysis, the YMC was simplified by assuming a linear progression and focusing solely on two specific transitions: from RC to OX and from OX to RB. For each of the three omics datasets (RNA-seq, ChIP-seq H3K18ac, ChIP-seq H3K9ac), a list comprising the log2FC and p-adj of each transition for all genes was provided as input for MAMBA.

In addition to gene-centric information, MAMBA incorporates metabolomic data to enhance the accuracy of flux predictions. However, it should be noted that certain experimentally measured metabolites, namely citrulline, Adenosine Diphosphate Ribose (ADPR) and nicotinic acid mononucleotide (NAMN), could not be included in the analysis, as they were not present in the pre-existing yeast metabolic model. As a result, log2FC and p-adj for only 47 metabolites were used for further analysis. The predicted fluxes were visualized using the Central Carbon Metabolism metabolic map of *Saccharomyces cerevisiae* in ESCHER (King et al., 2015).

3.5.2 Reduced Cost Analysis

Linear optimization problems may have multiple solutions with a maximum value of the objective function. MAMBA returns only one out of these many possible optimal solutions. Consequently, a sensitivity analysis was required to evaluate the significance of the predicted variables. For this purpose, binary variables, i.e. gene states (either active or inactive), were fixed to the value that they take in the solution returned by MAMBA. The problem was solved again using the simplex algorithm to obtain the fluxes and reduced costs (RC) of the remaining variables (reactions). The RC of the reaction accounts for the impact of each reaction flux (equal or different from zero) on the value of the objective function. Specifically, it is zero if increasing the variable (turning a reaction from inactive to active) does not penalize the objective function. Similarly, it is non-zero if increasing the variable has a negative impact on the objective function. That is, the reaction being active cannot be part of the optimal solution.

Subsequently, the reaction flux (V) and its associated RC were utilized to categorize reactions into three distinct groups: active reactions (V $>$ 0, RC = 0), potentially active reactions (V = 0 but RC = 0) and inactive reactions (V = 0, RC 0). For the purpose of downstream functional analysis, active reactions and potentially active reactions were combined and labelled as "active." Additionally, only those reactions where an increase in the corresponding variable penalized the objective function (V=0, RC != 0) were considered "inactive".

Lastly, Differentially Active Reactions (DARs) were defined as reactions being exclusively active in one of the conditions.

3.5.3 Pathway Enrichment Score

To explore the biological significance of these DARs, a pathway-level analysis using yeast pathways from KEGG database was conducted in RStudio.

Firstly, each DAR was associated with its corresponding KEGG ID, whenever possible. Secondly, each KEGG ID was assigned to one or multiple cellular pathways. Then, the total number of reactions was determined for each of the pathways. Finally, the Pathway Enrichment Score (PES) was defined as the percentage of active reactions for each pathway (Ugidos Guerrero, 2023). Given the metabolic pathway P containing n reactions, PES was calculated as the number of DARs associated to that pathway divided by total number of reactions (Eq. 2)

$$PES = 100 \times \left(\frac{\text{DARs of pathway P}}{\text{total number of reactions of pathway P}} \right) \tag{2}$$

The PES in the general KEGG pathway, calculated as the number of DAR (for all pathways) divided by the total number of reactions (of all pathways), was set as a relevance threshold. Those pathways having a PES higher than the general PES, were considered biologically relevant.

3.6 Gene Set Variation Analysis and Clustering

Gene Set Variation Analysis (Hänzelmann et al., 2013) was applied to the gene counts matrix of those genes involved in the pathways previously identified as biologically relevant. The analysis was conducted using GSVA package in RStudio, separately for the relevant pathways identified by MAMBA and FE methods. The resulting sample-wise GSVA scores were used for unsupervised clustering analysis.

4 Results and discussion

4.1 Omic data pre-processing

The obtention of RNA-seq gene counts involved several steps. Initially, raw reads data was acquired from the SRA database, followed by quality assessment and trimming procedures. The resulting reads were subsequently mapped against the reference genome of *Saccharomyces cerevisiae*.

The detailed process is exemplified for a single sample, which can be considered representative of the entire dataset. The raw reads obtained from the SRA database encompassed a total of 11 million reads. Prior to further analysis, a quality check was conducted using the fastqc tool. The examination revealed high quality scores both per base and per sequencing, indicating that the overall quality of the reads was satisfactory (Fig. 8).

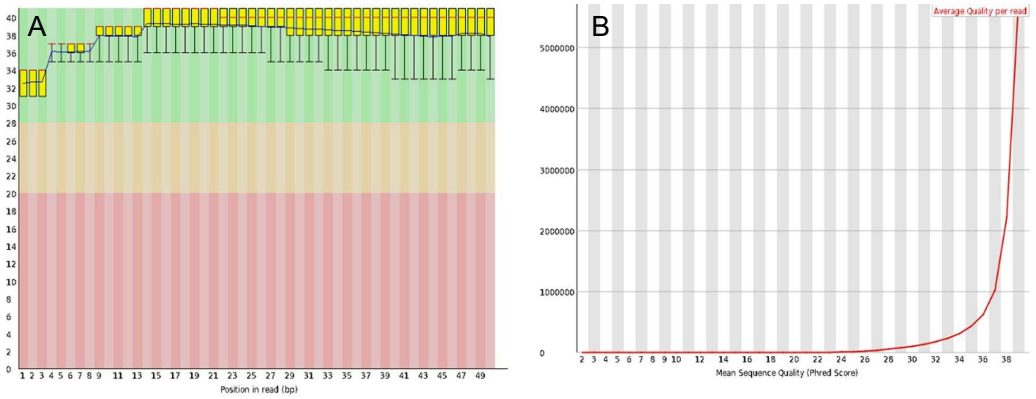


Figure 8: Quality assessment of RNA-seq reads using FastQC. (A) Per base sequence quality. (B) Per sequence quality scores. Statistics refer to a single sample, but are representative of the rest.

Nevertheless, the per base N quality exhibited a relatively elevated level towards the end of the reads (Fig. 9). Consequently, a trimming step was employed to eliminate the N content utilizing cutadapt. Notably, all reads were retained throughout this trimming process.

Following the mapping against the reference genome using the STAR, gene counts were calculated for unstranded RNA-seq data based on the characteristics of the library preparation. This analysis yielded the following results: 30,000 reads remained unmapped, 1 million reads were found to have multiple mapping locations (multimapped), 700,000 reads were mapped but did not correspond to any specific genomic feature, and 19,000 reads were ambiguously mapped. Ultimately, a total of 9 million reads were uniquely and unambiguously mapped to one gene in the reference genome and used for downstream analysis. The distribution of mapped reads across all samples can be observed in Figure 10.

4.2 Study of variability

4.2.1 Coefficient of variation

MAMBA operates on differential values between two comparing conditions, in this case the phases of the Yeast Metabolic Cycle (YMC): RC, OX and RB. To ensure the significance of the gene expression levels resulting from the differential expression analysis, replicates are needed.

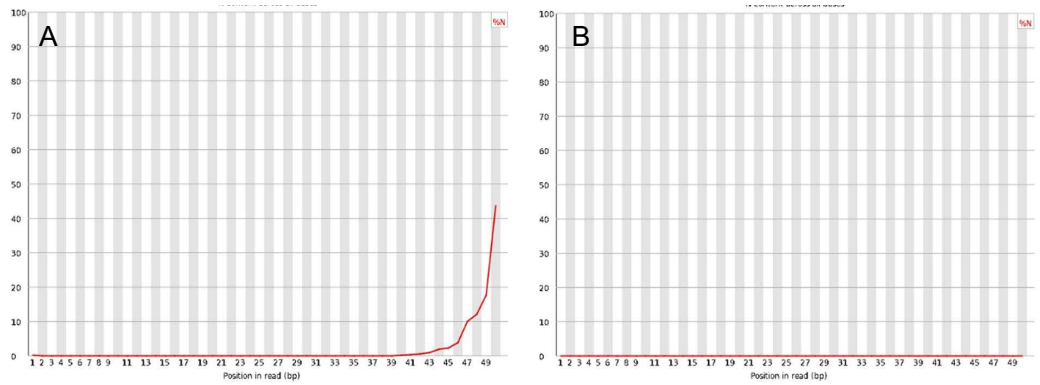


Figure 9: Quality assessment of RNA-seq reads using FastQC. (A) Per base N content of the original reads. (B) N content after trimming with cutadapt.

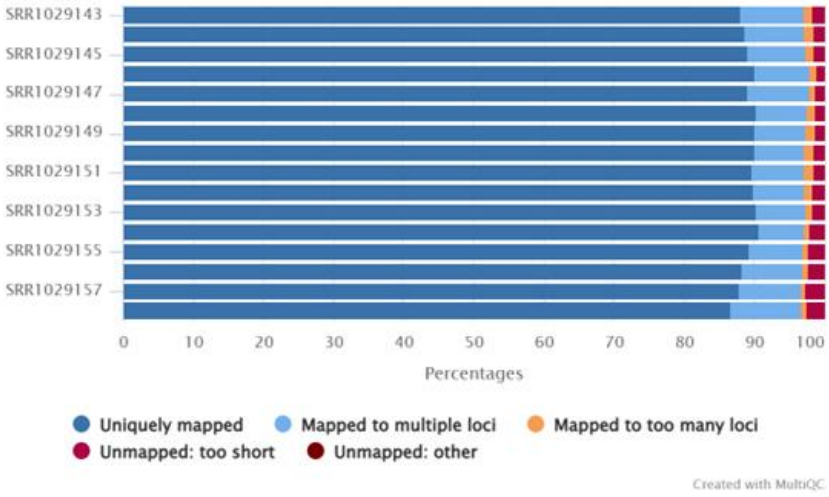


Figure 10: STAR Alignment Scores. Percentage of multimapped, unmapped and uniquely mapped reads of the 16 RNA-seq samples.

The original dataset comprised 16 different time points distributed along the cycle. In the study by Kuang *et al.* (2013), oxygen levels were used to categorize these time points into the corresponding metabolic phases. However, the samples are not replicates, as they had not been collected simultaneously.

Therefore, the first assumption to validate was whether time points within the same phase were sufficiently similar to be treated as biological replicates. This assumption holds true if the gene expression values of time points within the same phase exhibit greater similarity compared to those of time points belonging to different phases. In other words, the intra-phase variability should be significantly smaller than the inter-phase variability, or at least smaller than any other random grouping.

To evaluate gene expression variability within the phases, the distribution of variation coefficients (CV) for genes was analyzed (Table 1). OX phase, despite having a smaller sample size (3), exhibited the highest variability, as indicated by a median CV of 0.28. In contrast, the RB and RC phase, with sample sizes of 7 and 6 time samples, respectively, demonstrated significantly lower variability, with a median variation coefficient of 0.19.

To further validate the original biological-based grouping, the variation following original phase division was compared to the variation of randomly grouped samples. For this purpose, 50 random groupings were generated by randomly assigning 3, 7 and 6 samples to the OX, RB and RC phase, respectively. The CV mean of this new groups was computed for each phase and compared to the original one. Despite of the CV mean of these random groups being consistently

Table 1: Distribution of Coefficients of Variation (CV) in the three phases of the cycle. 3, 7 and 6 samples were assigned to OX, RB and RC, respectively, following the oxygen-based biological division.

Phase	min	Q1	median	Q3	max	sd
OX	0.0008228417	0.1622016	0.2761438	0.4473455	1.732051	0.3389449
RB	0.01718768	0.1190627	0.1919416	0.3304646	2.588891	0.2674267
RC	0.01925737	0.1212407	0.194726	0.3323307	1.361343	0.2606483

higher for all phases compared to the original one (Fig. 11), certain random groupings exhibited lower variability than the original grouping for specific phases, particularly in the case of OX (Table 2).

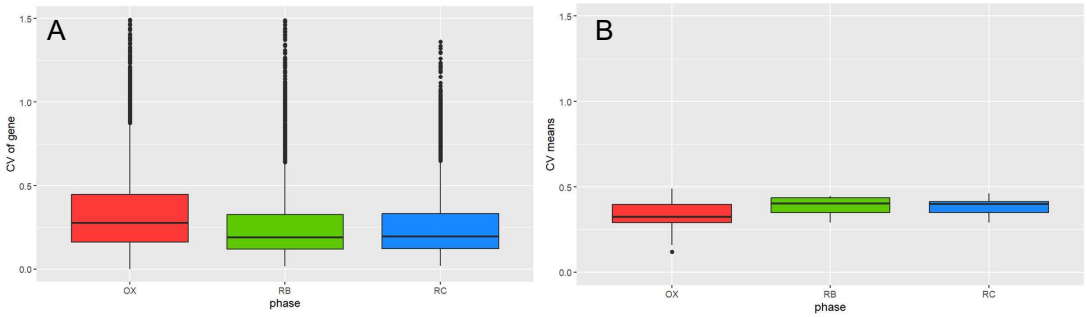


Figure 11: Coefficient of Varitation (CV) distribution. (A) CV of each gene for OX, RB and RC following the original biological-based phase division. (B) Mean of the CV of all genes in each phase for 50 randomly generated groupings maintaining sample size.

However, evaluating the phases independently does not provide sufficient information because a random grouping that exhibits lower variability in one phase but higher variability in the other two should not be considered superior. In other words, for a grouping to be deemed more suitable than the original one, the average of CV means for all three phases must be smaller. The mean of the three phases was calculated for each random group and compared to the original one (Fig. 12).

Table 2: Coefficient of Variation (CV) means distribution. 50 groupings were randomly generated mantaining the original sample size.

Phase	min	Q1	median	Q3	max	sd
OX	0.1188018	0.290143	0.3238123	0.3955575	0.4901735	0.3330523
RB	0.2908184	0.3483872	0.403333	0.4350434	0.447873	0.3919544
RC	0.2895896	0.3493645	0.39949	0.4131738	0.4605378	0.3821736

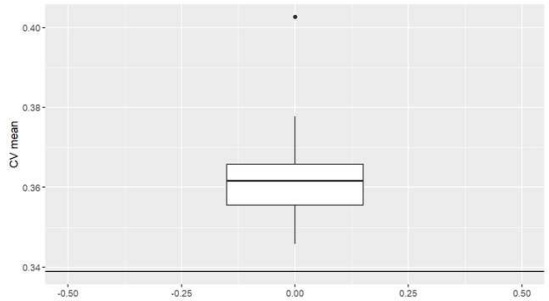


Figure 12: Coefficient of Variation (CV) means distribution. Distribution of CV means of the three phases for each random grouping. The CV mean following the original grouping is presented as a base line.

Following the biological-based phase division, the CV mean was 0.339, while the mean for the

randomly generated groups was consistently higher, between 0.347 and 0.378 in all simulations. Based on these results, it can be inferred that the initial oxygen-based grouping is optimal, as it demonstrated the lowest intra-phase variability compared to any alternative grouping.

4.2.2 Clustering

Principal Component Analysis (PCA) is a statistical technique used to analyze and visualize the variability in a dataset. It transforms the original variables into a new set of uncorrelated variables, called principal components (PC), which capture the maximum amount of variance in the data. A 2D plot of the PC allows to visually identify clusters of closely related data points, providing insights into the patterns and relationships between the samples (Jolliffe & Cadima, 2016).

In the case of RNA-seq analysis (Fig. 13), it was found that approximately 60.44% of the variance could be explained by the first two principal components. PC1 effectively distinguished the RB and OX from RC phase, while PC2 primarily separated RB from OX. This suggests that RB and OX phases exhibit greater similarity to each other compared to RC, which stands out as the phase with the most distinct gene expression patterns. These findings align with the observations from the expression-based unsupervised clustering (Fig. 14), where RB and OX are clustered together and separated from RC, providing further consistency to the overall analysis.

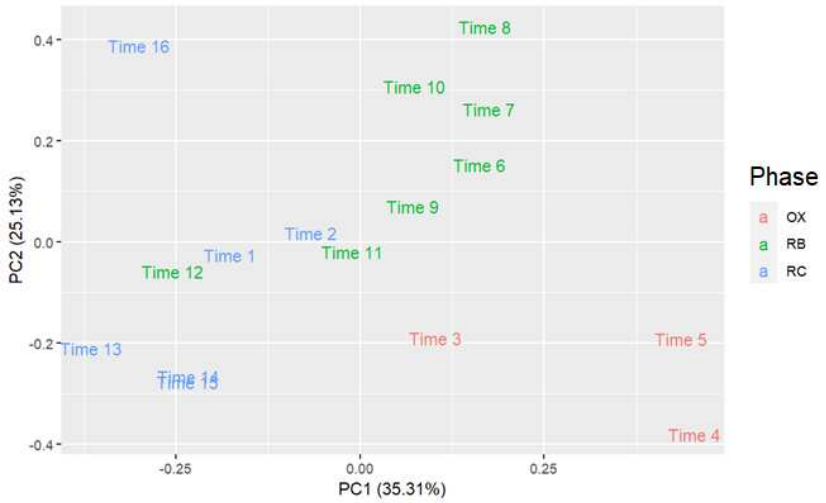


Figure 13: PCA of RNA-seq samples. Time samples are coloured according to the original phase division.

It is noteworthy that Time 12 is clustered within RC samples. This, together with the fact that oxygen levels at Time 12 were more similar to RC than RB (Fig. 7) suggest that this particular sample is not representative of the gene expression profile of its phase. Given that the RB phase has the largest sample size, the omission of Time 12 from the dataset for downstream analysis is possible without compromising the significance of the DE analysis. As evident from the heatmap (Fig. 14), a similar situation occurs with Time 3, which is clustered within the RC samples. However, in the case of Time 3, it cannot be excluded from the analysis due to the limited sample size of the OX phase.

In the case of ChIP-seq analysis, the clustering of samples before H3 normalization (Supp. Fig. 33) and after H3 normalization (Fig. 15) demonstrated consistent alignment with the phase division for both histone modifications. This means that samples belonging to the same phase exhibited closer proximity to each other and were more distinct from samples in other phases. The observed clustering patterns validate the effectiveness of H3 normalization in preserving the inherent phase-specific characteristics of the histone modification data.

It is worth noting that in the case of K18ac analysis, Time 11 exhibited a closer proximity

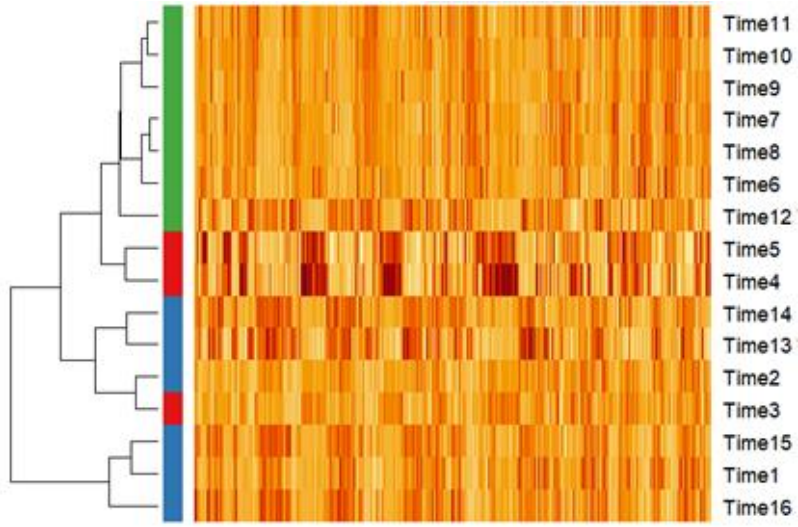


Figure 14: Unsupervised clustering of the samples based on RNA-seq gene expression values. Columns refer to genes and rows to time points. Side colors represent the YMC phase associated to each time point (blue for RC, green for RB, red for OX).

to the RC samples compared to the RB samples. This observation aligns with the anticipated pattern based on the similarity of oxygen levels (Fig. 7), which indicate a greater resemblance to the RC phase. However, despite this observation, Time 11 was not excluded from the analysis in order to maintain consistency in the sample size across the ChIP-seq datasets.

Finally, the PCA revealed that the variability in the ChIP-seq data exhibited a circular pattern. In addition to grouping samples according to their respective YMC phases, the PCA analysis captured the temporal differences between samples within the same phase.

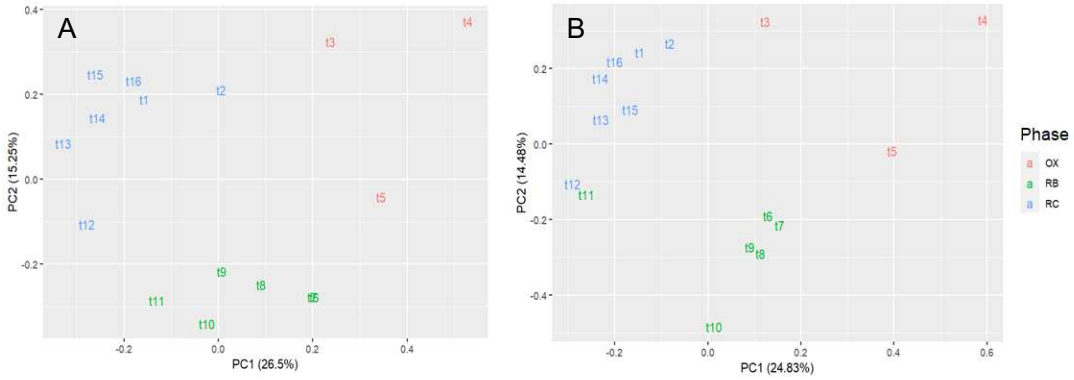


Figure 15: PCA of ChIP-seq samples. (A) H3K9ac after H3 normalization (B) H3K18ac after H3 normalization.

Regarding the analysis of metabolomics data (Fig. 16), PC1 clearly separates the RB phase from the OX and RC phase, indicating a distinct metabolic profile between these phases. On the other hand, PC2 differentiates the RC phase from the OX phase, suggesting that RC and OX exhibit a greater similarity in terms of metabolic profiles compared to the RB phase. This contrast in the metabolomic data's pattern of separation is different to the epigenetic and transcriptomic data, where RB and OX were found to be more similar.

Furthermore, similar to ChIP-seq data, the metabolomic data also demonstrates YMC temporal variability. The positioning of samples along the principal components suggests a temporal progression of metabolite profiles during the YMC, where changes in metabolic states are observed between different phases and gradual variations occur within each phase. This temporal variability observed in the metabolomic dataset further emphasizes the dynamic nature of the metabolic processes underlying the YMC.

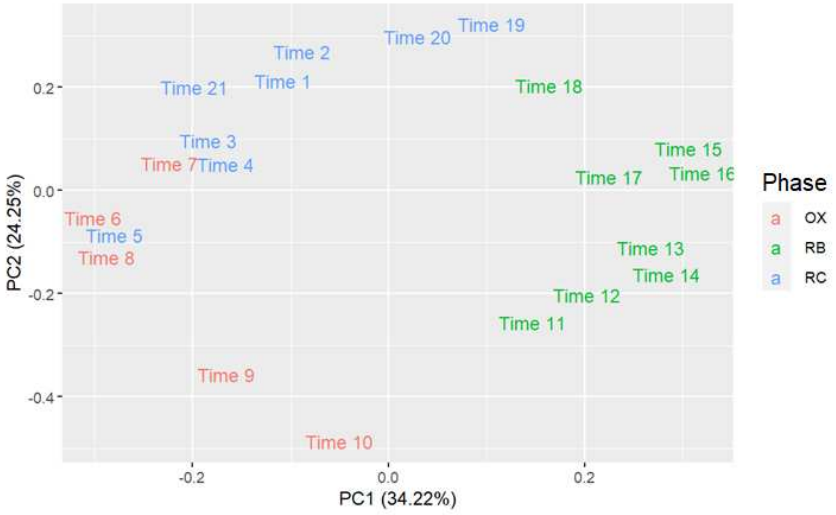


Figure 16: PCA of metabolomic samples. Metabolomic information includes 21 samples distributed along the 3 phases of the cycle.

In summary, the final datasets used for downstream Differential Expression (DE) analysis included 15 time samples for transcriptomic data (excluding Time 12), 16 time samples for both histone modifications, and 21 time samples for metabolomic data. The samples were categorized into three phases based on the original oxygen-based division and were treated as biological replicates for downstream analysis.

4.3 Differential Expression analysis

The general goal of the DE analysis is to determine which genes are expressed at different levels between conditions, as they can offer biological insight into the processes affected by the condition(s) of interest. In this work, results of DE analysis were introduced in MAMBA as input to define genes and metabolites constraints for metabolic modelling.

In the case of RNA-seq data (Table 3), nearly half of the genes had a p-value below 0.05 for all transitions, confirming distinct functional profiles across the different phases. By applying an additional magnitude criterion of absolute log2 Fold Change (log2FC) > 2, the list of significantly differentially expressed (DE) genes is substantially reduced. Notably, the transition from RC to OX exhibits the highest number of DE genes, indicating pronounced differences between these phases, consistent with the findings from the exploratory variance analysis.

Table 3: Differentially Expressed (DE) genes. Number of DE genes for the 3 transitions based on RNA-seq data.

	p-adj <0.05	p-adj <0.05 and |log2FC| >2
From RC to OX	3818	592
From OX to RB	2763	311
From RB to RC	3172	277

Conversely, the transitions entering and leaving RB (from OX to RB and from RB to RC) exhibit a lower number of DE genes, implying that RB may serve as an intermediate stage between OX and RC. This observation is further supported by the overlap analysis of DE genes between different transitions, where the transitions from OX to RB and RB to RC exhibit the least overlap (Fig. 17). This suggests that genes that become activated or inactivated upon entering RB are unlikely to undergo substantial changes when this phase transitions to the subsequent phase.

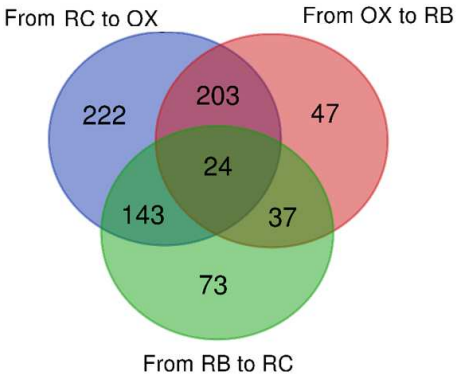


Figure 17: Overlap between the Differentially Expressed (DE) genes in the different transitions based on RNA-seq. The intersection between the circles represent the number of DE genes identified in the two or three corresponding transitions.

The DE analysis of epigenetic data reveals a significant number of genes with a p-value below 0.05 as well (Table 4). However, when considering the log2FC threshold, there are notably fewer genes meeting the criteria, particularly in the case of H3K18ac (Table 4). This observed discrepancy can be attributed to the distinct characteristics of omics technologies employed, which result in a narrower dynamic range in ChIP-seq compared to RNA-seq. While gene expression levels can exhibit considerable variation, chromatin accessibility as measured by ChIP-seq is confined to a single genome per cell, leading to a more constrained range of observed changes.

Table 4: Differentially Expressed (DE) genes. Number of DE genes for the 3 transitions based on H3K18ac ChIP-seq data.

	p-adj <0.05	p-adj<0.05 and |log2FC| >2
From RC to OX	3494	49
From OX to RB	2230	5
From RB to RC	2243	30

Comparing the two histone modifications (Fig. 18A), there is a substantial overlap in the set of DE genes, indicating that they likely to be cooperating in the regulation of shared biological processes. This may imply that histone marks have a combined effect that may not be fully captured when analysing them separately.

However, when comparing the DE genes from transcriptomic and epigenetic data, it becomes apparent that the majority of DE genes are exclusive to either one of the omics, with very few genes overlapping between the two (Fig. 18B).

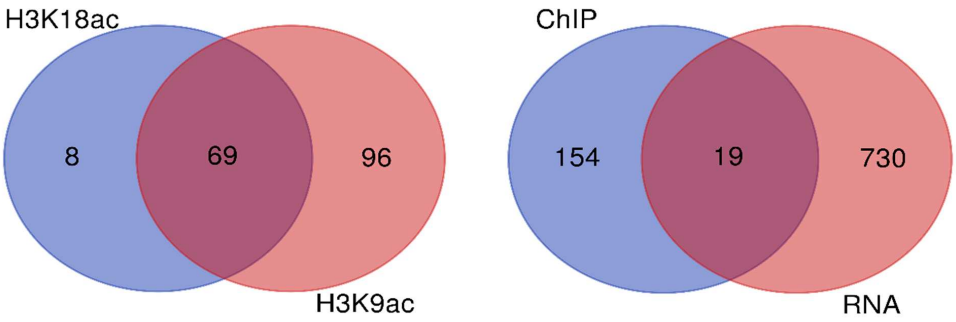


Figure 18: Overlap of Differentially Expressed (DE) genes altogether for the three transitions (A) Overlap between histone modifications. (B) Overlap between epigenetic and transcriptomic data. The intersection between the circles represents the number of DE genes identified by the two omics.

The observation that transcriptomic and epigenetic data provide distinct information about gene expression regulation is indeed significant. It highlights the complementary nature of these two types of data and suggests that epigenetic marks captured by ChIP-seq can provide valuable insights into the regulation of metabolic pathways through epigenetic mechanisms.

4.4 Functional Enrichment Analysis

To gain greater biological insight on the DE genes, several analyses were performed, including Gene Set Enrichment Analysis (GSEA) and Over Representation Analysis (ORA). Both methods aim to identify biological functions that are enriched in the set of DE genes, but approaches and statistical models vary substantially.

4.4.1 Gene Set Enrichment Analysis

The DE analysis highlighted the challenge of comparing gene-associated p-values and log2FC between omics technologies due to the different dynamic ranges of RNA-seq and ChIP-seq. This disparity renders enrichment analysis tools that rely on an arbitrary pre-selection of "interesting" genes based on significance or magnitude thresholds, such as ORA, less suitable for this scenario. For this reason, we decided to use GSEA, which circumvents the need for pre-selection by taking as input a ranked list of all genes based on a statistical measure, such as the Walt statistic. By doing this, GSEA provides a more comprehensive and unbiased assessment of biological functions associated to the DE genes, overcoming the limitations posed by comparing p-values and log2FC directly.

Functional Enrichment (FE) analysis using GSEA allowed for the identification of pathways that were upregulated or downregulated during the three phase transitions within the YMC. Specifically, during the transition from RC to OX phase (Fig. 19), there is a consensus among the three omics regarding the activation of pathways related to protein synthesis. These pathways mainly include processes associated with transcription and translation. The first category encompasses cellular functions involved in nucleotide availability such as purine metabolism and pyrimidine metabolism, as well as RNA polymerization pathways including RNA polymerase, basal transcription factors, spliceosome and RNA transport. The second category comprises pathways related to ribosome function, including structural genes involved in ribosomal RNA (rRNA) and genes associated with ribosome assembly and biogenesis.

It is noteworthy that functional characterization based on ChIP-seq data does not reveal pathways related to amino acid availability, in contrast to the clear identification of these functions in RNA-seq. These pathways include the biosynthesis of amino acids (lysine, valine, leucine, isoleucine, methionine, cysteine and aromatic aminoacids). Other more isolated pathways include DNA replication, whose activation is further emphasize in the next transition (OX to RB).

In contrast, when examining the pathways that are inactivated during the RC to OX transition, it becomes apparent that energy production-related pathways are significantly affected. These include sugar metabolism pathways such as galactose, starch and sucrose, fructose and mannose metabolism, as well as oxidative respiration-related pathways encompassing glycolysis, pyruvate metabolism, citrate cycle, and oxidative phosphorylation. Additionally, pathways associated with the mobilization of alternative energetic compounds, such as glycerolipid metabolism, fatty acid degradation, amino sugar and nucleotide sugar metabolism, and propanoate metabolism, are also observed to be inactivated. These findings suggest that energy production is a key process during the RC phase of the YMC. However, the interpretation of the other two transitions (entering and leaving RB) is not as straight-forward. From OX to RB, there is a consensus among the omics data regarding the inactivation of protein synthesis-related pathways (Fig. 20). This indicates that these pathways are overregulated in the OX phase and become downregulated as

the YMC transitions into RB.

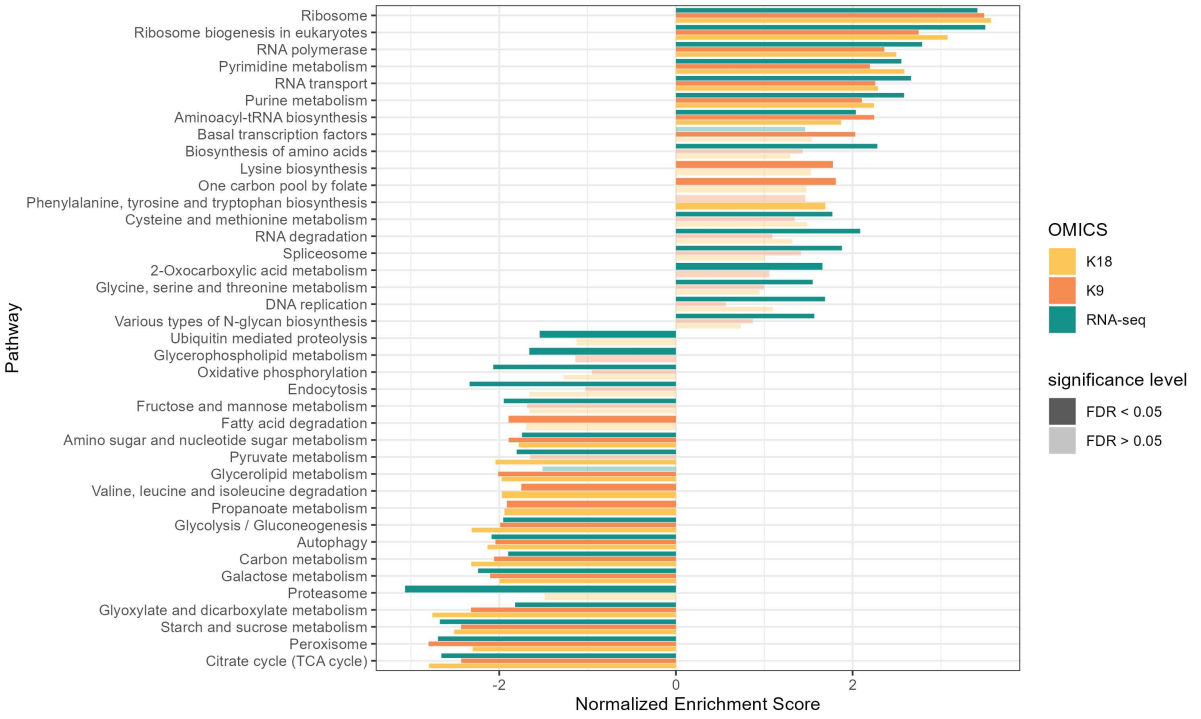


Figure 19: GSEA results OX vs RC. Pathways with differential activity in the RC to OX transition based on RNA-seq and ChIP-seq. Pathways with a positive enrichment score (ES) are activated in the transition, while pathways that have a negative ES are inactivated.

In terms of upregulated pathways, there is a significant overlap between OX to RB and RB to RC transitions. This suggests that RB serves as a transitional phase between OX and RC, exhibiting similarities to both phases. The upregulated pathways primarily pertain to energy production, further highlighting the importance of energy metabolism. Interestingly, the RNA-seq analysis reveals the activation of DNA replication-related processes, aligning with the observations made in the RC to OX transition. These include pathways involved in mismatch repair and homologous recombination, underscoring the crucial role of DNA replication during RB and implying a potential connection with cell division.

GSEA provides a broad understanding of the progression of the YMC, but its focus on transitions makes it difficult to achieve a comprehensive and accurate functional profiling of the three phases independently. Nonetheless, several key points can be inferred from the analysis.

Firstly, protein synthesis is activated upon entering the OX phase, indicating the initiation of cellular resource generation to support metabolic activity and cell division. Secondly, DNA replication-related processes are observed in the RB phase, suggesting ongoing DNA replication during this stage. Lastly, the RC phase exhibits the highest levels of energy production, implying a shift towards energy extraction from lipid reservoirs. These findings align with previous knowledge of the YMC, as described by Casaní Galdón (2021). According to this knowledge, the cycle begins with protein synthesis to provide the necessary cellular resources for metabolic activity and cell division. This is followed by a phase of high energy consumption, primarily through respiratory metabolism, to facilitate cell division. Once cell division is completed, the cycle enters a phase where energy is derived from lipid reservoirs until enough energy is accumulated to initiate the cycle again.

Regarding the performance of both omics, RNA-seq outperforms ChIP-seq in terms of sensitivity, as it identifies a higher number of relevant pathways in all three transitions. However, ChIP-seq data still capture similar cellular functions, and there are even exclusive pathways identified by ChIP-seq, such as ABC transporters. Additionally, there appears to be a slight

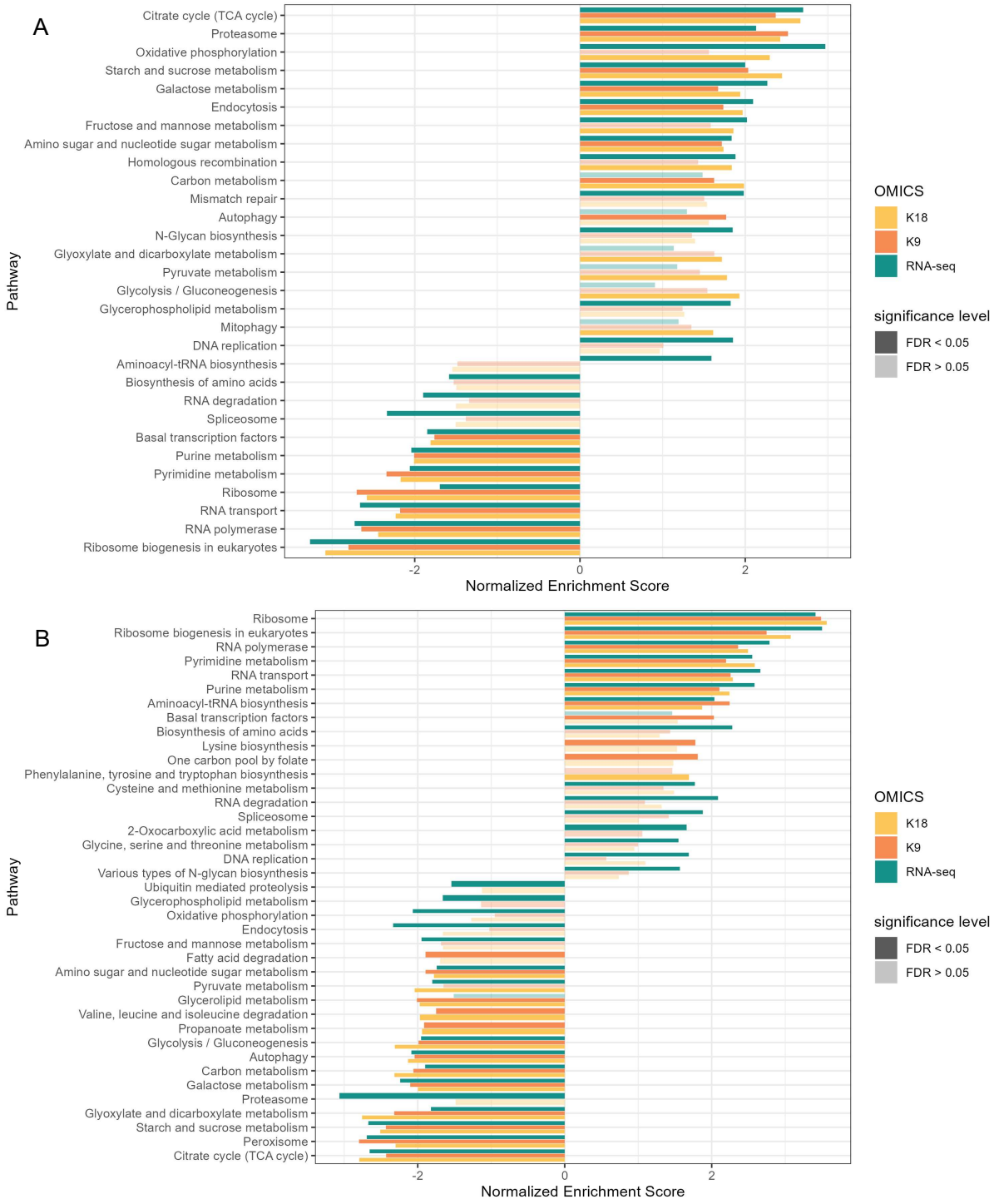


Figure 20: GSEA results for 2 transitions. Pathways with differential activity based on RNA-seq and ChIP-seq in the (A) OX to RB transition and (B) RB to RC transition. Pathways with a positive enrichment score (ES) are activated in the transition, while pathways that have a negative ES are inactivated.

delay between epigenetic and transcriptomic data, as some processes activated in one transition based on RNA-seq may be activated in a different transition based on ChIP-seq.

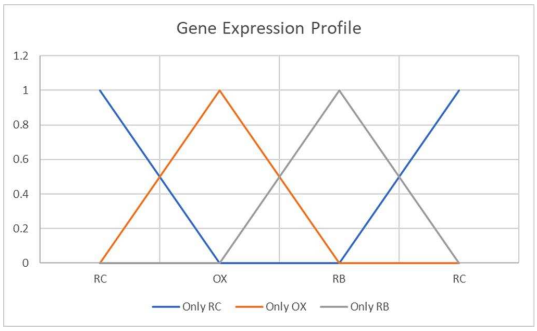
Comparing the two histone modifications, H3K9ac and H3K18ac, there is no clear evidence of one being superior to the other. Each modification seems to better recapitulate functional differences in specific transitions. For example, H3K9ac captures differences in the RB to RC and RC to OX transitions, while H3K18ac highlights differences in the RB to OX transition. This suggests that the histone modifications may have distinct roles in different phases of the

cycle or that they have a coordinated effect that cannot be fully understood through independent analysis.

4.4.2 Over Representation Analysis

ORA is a statistical method that determines whether genes from pre-defined sets (those belonging to a specific KEGG pathway) are present more than expected (overrepresented) in a subset of your data (significantly under- or overexpressed genes). While the focus of GSEA was to study phase transition (which genes are up- or downregulated when going from one phase to the next one), ORA's focus was set in the characterization of functional differences for each phase independently. In other words, the goal of this analysis was to identify which biological functions are differentially and exclusively active in each of the phases. Genes related to these biological functions are expected to be upregulated when the cell enters that phase (log2FC $>$ 0) and downregulated (log2FC $<$ 0) when it enters the next one or, in other words, upregulated in that phase with respect to the other two (Fig. 21).

A



B

	RC to OX	OX to RB	RB to RC
Only in OX	↑	↓	
Only in RB		↑	↓
Only in RC	↓		↑

Figure 21: Differentially Expressed genes for ORA. (A) Simplified representation of the gene expression profiles that are selected for ORA (B) In order to identify metabolic functions characteristic of each YMC phase, only those genes being exclusively active between conditions were considered.

ORA based on the genes that are exclusively active in the OX phase (Fig. 22), provided a clear and consistent observation of high protein synthesis. This is evident from the enrichment of processes related to transcription, such as RNA polymerase activity, transcription factors, and nucleotide metabolism. Similarly, pathways associated with translation, such as the ribosome and aminoacyl-tRNA biosynthesis, are also identified by all three omics datasets, and they exhibit similar levels of significance.

In the RB phase (Fig. 23), there is also a notable overlap in the functional pathways identified by the different omics datasets. Cell cycle is consistently identified by RNA-seq and ChIP-seq datasets, but it exhibits the lowest enrichment scores among the identified pathways.

Transcriptomic data highlights the activation of pathways related to O- and N-type biosynthesis, as well as cell division-related processes such as mismatch repair and DNA replication. These findings align with the previous results from GSEA connecting RB and cell division.

In contrast to OX, in which the overlap between epigenetic marks was almost total, H3K9ac and H3K18ac exhibit differences in the pathways identified as upregulated in the RB phase (Fig. 23). H3K9ac shows a strong association with the proteasome pathway, which is not observed in the RNA-seq or H3K18ac results. Both H3K9ac and H3K18ac show an association with oxidative phosphorylation, although it is not significant in the RNA-seq analysis. Notably, H3K18ac is strongly associated with the DNA replication machinery, with DNA replication, mismatch repair, and homologous recombination identified as the most significant pathways.

In the RC phase (Fig. 24), transcriptomic data highlights pathways associated with energy production from various resources, including lipids (fatty acid degradation, peroxisome, propanoate) and carbohydrates (starch and sucrose metabolism, glycolysis, pyruvate metabolism). This emphasizes the significance of energy production in the RC phase. Additionally, processes

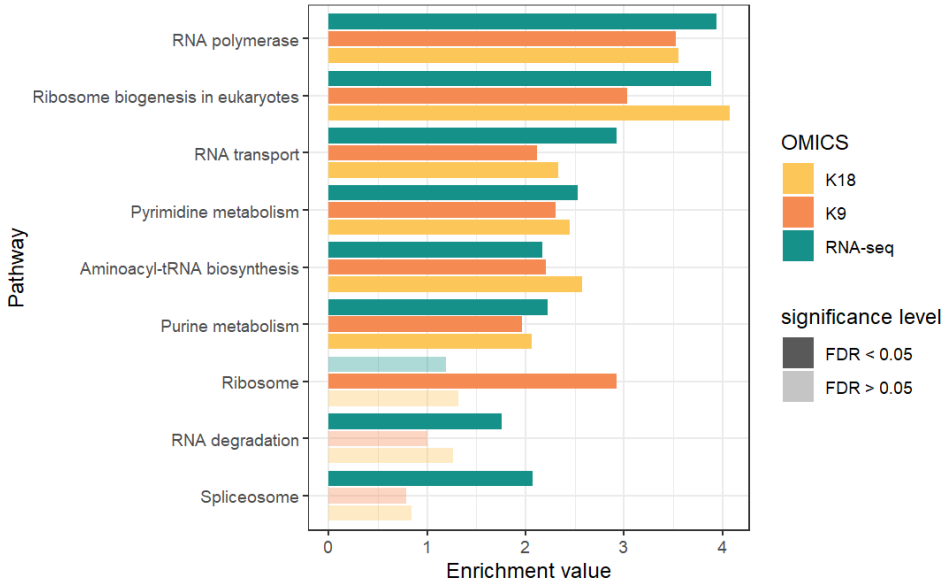


Figure 22: Over Representation Analysis results for OX. Pathways significantly enriched in DE genes based on RNA-seq data, H3K9ac data and H3K18ac data.

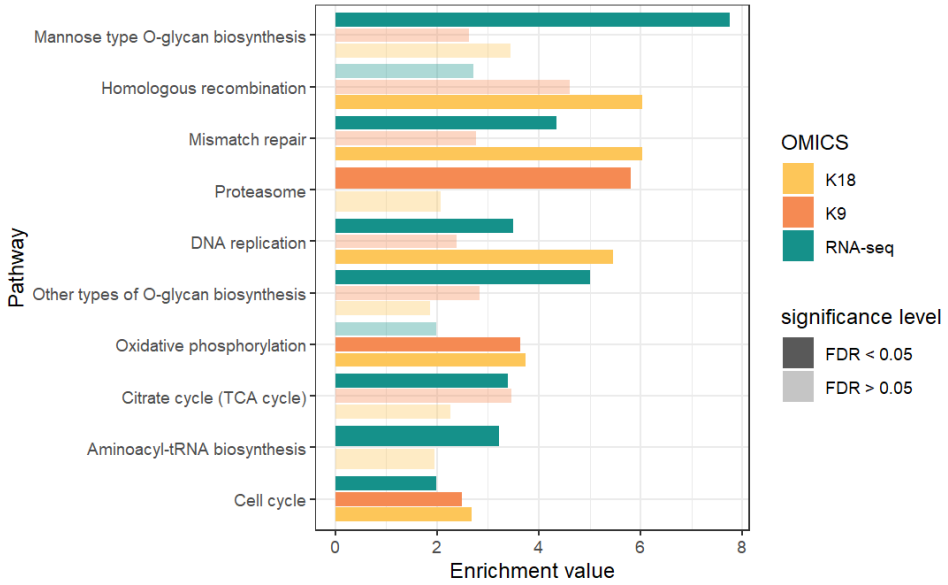


Figure 23: Over Representation Analysis results for RB. Pathways significantly enriched in DE genes based on RNA-seq data, H3K9ac data and H3K18ac data.

such as autophagy and endocytosis are also identified as active during this phase. An intriguing finding is the contrasting results regarding the activation of the proteasome pathway in the RC phase. While transcriptomic data shows a high enrichment score for the proteasome pathway, indicating its activation in this phase, neither of the histone modifications (H3K9ac and H3K18ac) exhibit any significant association with the proteasome pathway in the RC phase.

However, it is worth noting that the H3K9ac mark strongly anticipates the activation of the proteasome pathway in the RB phase, where it shows the same high enrichment score as observed in the RC phase transcriptomic data. This discrepancy implies that histones are modified to favour gene expression, but expression levels can change without altering chromatin accesibility. This suggest a possible time shift or a differential regulatory mechanism in the activation of the proteasome pathway between the RB and RC phases. Epigenetic data, in addition to aligning with the energetic metabolism pathways identified in RNA-seq, reveal the activation

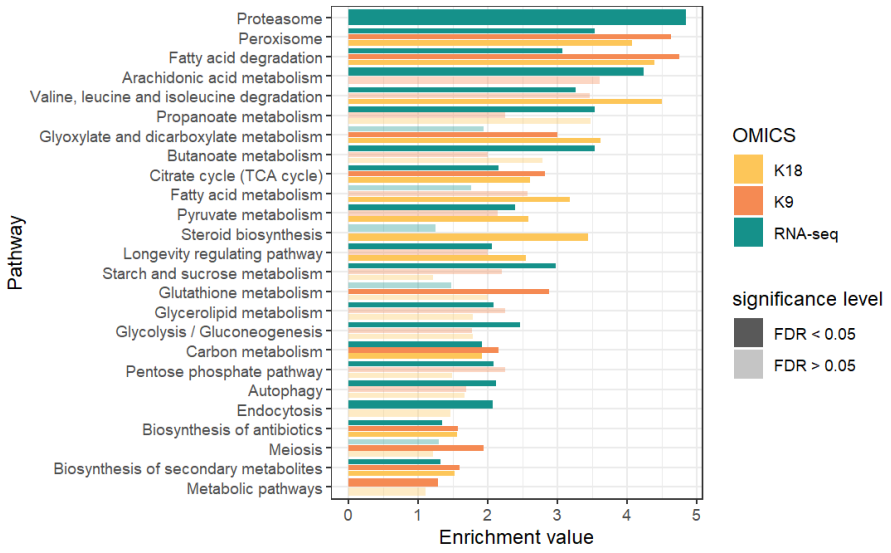


Figure 24: Over Representation Analysis results for RC. Pathways significantly enriched in DE genes based on RNA-seq data, H3K9ac data and H3K18ac data.

of pathways related to secondary metabolites, steroid biosynthesis, and antibiotics biosynthesis. These pathways were not identified in the GSEA analysis, indicating the unique contribution of epigenetic information in capturing specific biological functions.

Altogether, the integration of ORA results from epigenetic and transcriptomic datasets has provided a more comprehensive and precise functional profiling of the three phases of the YMC. Briefly, the OX phase is characterized by high protein synthesis, RB is associated with oxidative phosphorylation and cell division, and RC focuses on energetic production from lipids and carbohydrates. The YMC begins with protein synthesis, transitions to respiratory metabolism for cell division, and concludes with energy extraction from lipid reservoirs. Notably, a biological delay between transcriptomic regulation and epigenetic modifications could be observed. Processes like proteasome, strongly highlighted by ChIP-seq in RB, appear in RC in the RNA-seq results, which may indicate that changes in histone modifications are translated into gene expression changes at a later time point.

Unlike the GSEA analysis, which showed lower sensitivity for ChIP-seq data, performing ORA allowed for meaningful insights from both epigenetic and transcriptomic data, as it focus on significance and direction of the change rather than the magnitude. However, the different dynamic range of these datasets underscores the need for more adapted analysis procedures for both omics. Furthermore, the differences in upregulated pathways observed between omics datasets suggest that each dataset provides unique and complementary information about the YMC. While there is some overlap, the presence of exclusive pathways in specific omics indicates that integrating these findings is crucial for capturing a more complete understanding of the molecular events during the different phases of the YMC.

4.5 MAMBA

4.5.1 Flux prediction

In order to model the metabolic activity of the YMC, we used MAMBA, a tool that allows for the incorporation of transcriptomics or epigenomics in combination with metabolomics data. The predicted fluxes, i.e. the activity of the reactions of the metabolic network, were visualized using ESCHER (King *et al.*, 2015). The advantage of this approach is that, unlike FE analysis, which is limited to an overview of pathway activity, the visualization of MAMBA's prediction in

ESCHER allows for the identification of concrete reactions that are differentially active in the different conditions.

For example, when examining the fluxes based on ChIP-seq data in the OX phase (Fig. 25), it was observed that the oxidative carboxylation of pyruvate (Equation 3) is inactive.

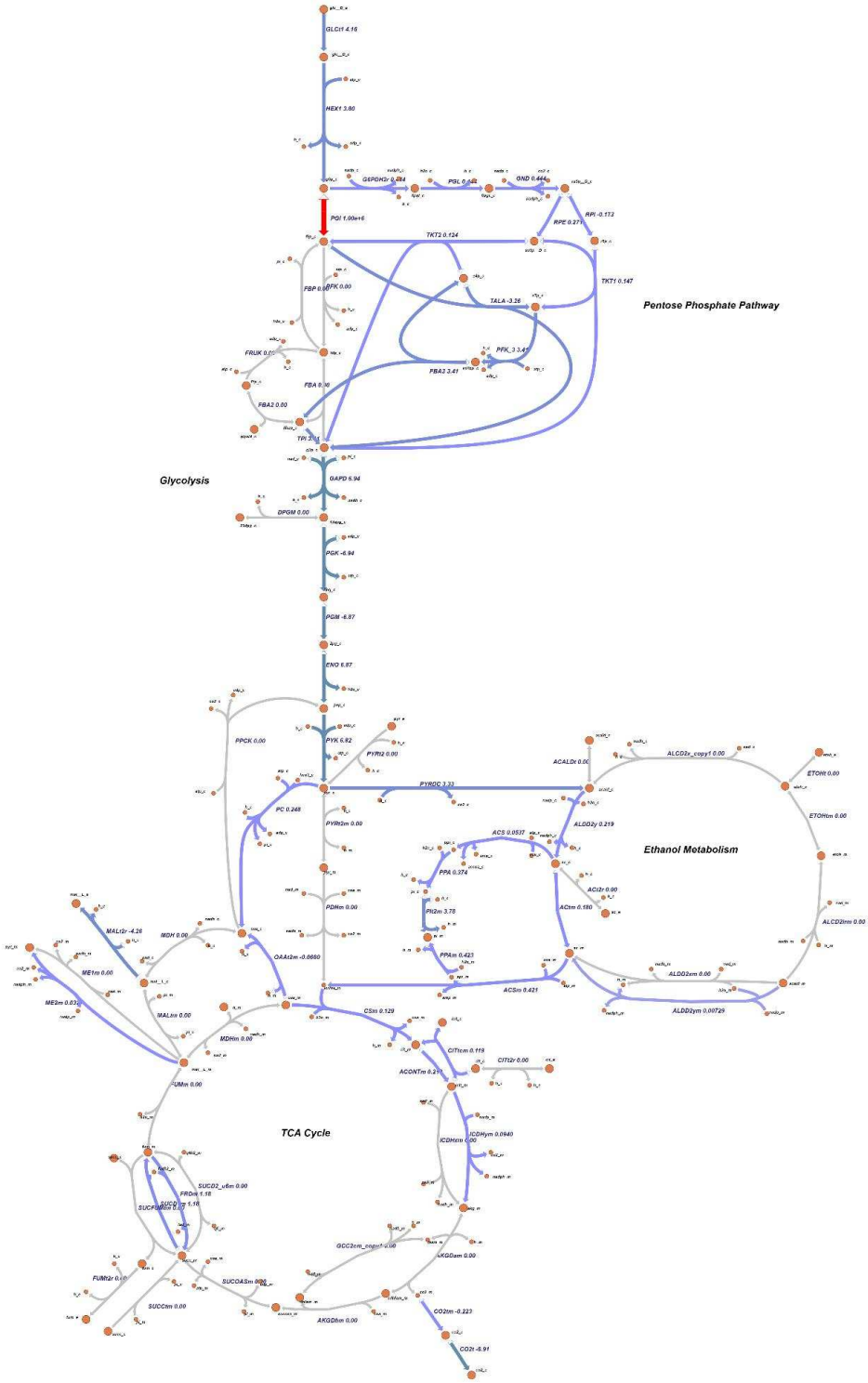


Figure 25: Metabolic map of Central Carbon Metabolism in *Saccharomyces cerevisiae* (iMM904). The intensity of the colors is proportional to the predicted flux based on H3K9ac information. Generated using ESCHER.

Cytoplasmic pyruvate (pyr_c) does not enter the mitochondria using pyruvate mitochondrial transport via proton symport but is instead being diverted towards ethanol metabolism and entering the mitochondria through an alternative route.

A big advantage of ESCHER is that it can also incorporate multiple sets of fluxes, allowing for comparisons between different conditions or omics data. By comparing the flux predictions based on RNA-seq for RC and OX phases, specific differences in pyruvate metabolism were also observed.

Specifically, the oxidative decarboxylation of pyruvate catalysed by the pyruvate dehydrogenase complex (Equation 3) was found to be significantly more active in the RC phase compared to OX (Fig. 26) This finding aligns with previous observations that indicated the importance of carbohydrate metabolism, including pyruvate metabolism, in the RC phase and indicates again an alternative metabolism in the case of OX.

$$\text{Pyruvate} + \text{CoA} + \text{NAD}^+ \rightarrow \text{acetyl-CoA} + \text{NADH} \tag{3}$$

However, it should be noted that this approach is limited by the reactions included in the map provided by ESCHER, which represents only a small part of the central carbon metabolism and does not encompass the entire metabolic network. This lack of biological context around specific reactions hinders the interpretation of many observed differences in reaction fluxes. Regarding the comparison of reaction activities between RNA-seq and ChIP-seq conditions, no significant differences were observed. The limited set of reactions in the map prevented the determination of specific differences in reaction activities between the two omics datasets.

To address these limitations and transform MAMBA's solution into biologically interpretable results, further analysis was performed to summarize the activity of individual reactions into specific pathways. This allowed for a more comprehensive understanding of the overall activity of pathways, providing a broader context for the functional characterization of the cycle.

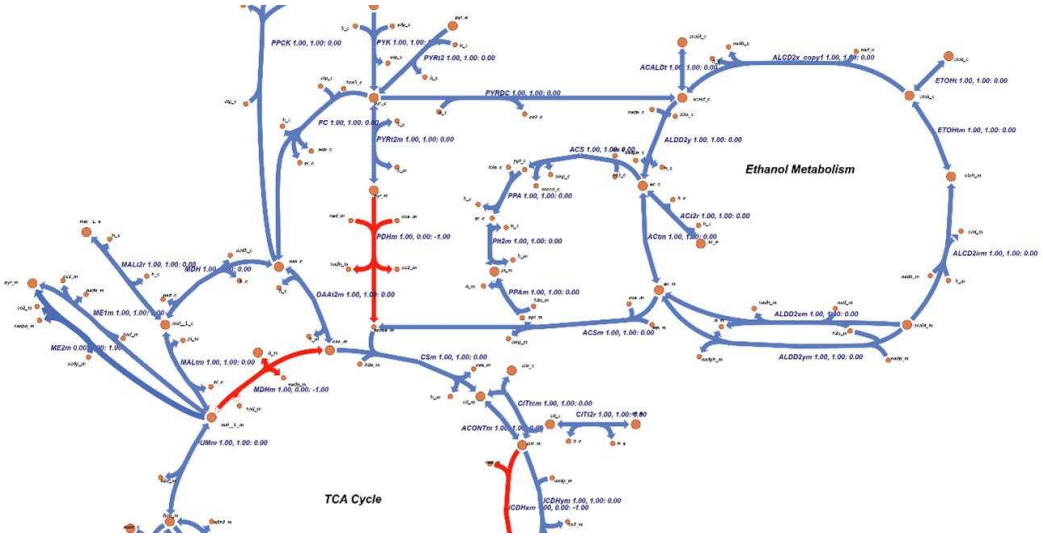


Figure 26: Partial metabolic map of Carbon Central Metabolism in *Saccharomyces cerevisiae*. Reactions highlighted in red represent those reactions that are inactivated when transitioning from RC to OX (significantly more active in RC than OX). Only reaction state (active / inactive), not the flux, was considered, based on RNA-seq data. Generated using ESCHER.

4.5.2 Pathway Enrichment Score

In order to identify pathways with differential activity between conditions, a similar approach to FE analysis was employed. Specifically, reactions that were exclusively active in each phase, meaning they had a non-zero flux value only in one of them, were designated as Differentially

Active Reactions (DARs). Subsequently, a Pathway Enrichment Score (PES) was calculated to as the percentage of DARs within each pathway.

To establish the association between reactions, pathways and genes, the KEGG pathway database was utilized. However, it is important to note that only a subset of reactions (625 out of 1577) could be annotated with the corresponding KEGG IDs, thereby constraining the analysis to the annotated reactions. The pathways identified and their corresponding PES values for each of the three phases and three omics datasets are provided as supplementary material (Supp. Fig. 34, 35, 36).

Upon initial examination, it is evident that the predictions based on ChIP-seq data identified an equal or larger number of differently activated pathways for all three conditions in comparison to RNA-seq (Fig 27). This contrasts with enrichment-based methods, which suggested that transcriptomic data exhibited higher sensitivity. However, the analysis conducted using MAMBA allowed for obtaining comparable information in terms of both the quantity and significance of pathways from both omics datasets. This difference in findings may be attributed to the fact that MAMBA's analysis of pathway activity is focused on reactions rather than genes. By transforming quantitative gene-centric omics data into binary gene activation states, MAMBA mitigates the impact of the distinct dynamic ranges observed in RNA-seq and ChIP-seq gene counts.

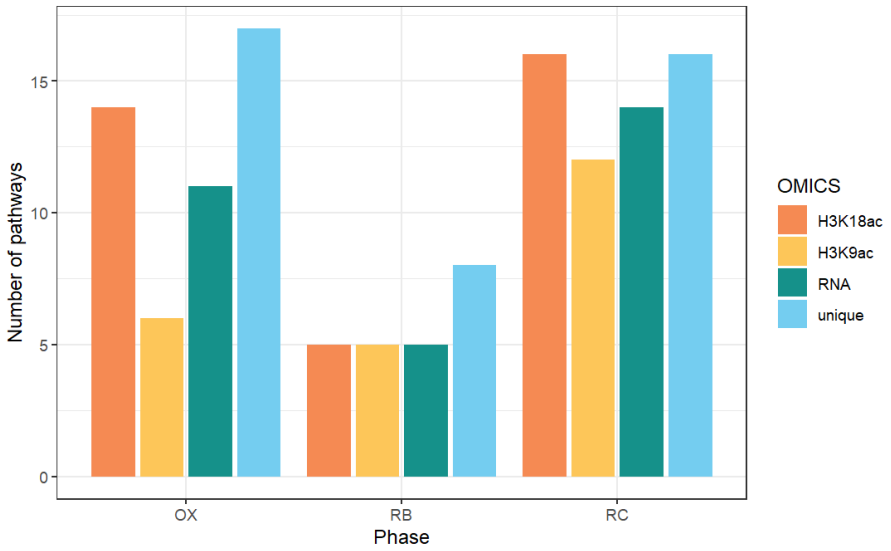


Figure 27: Number of differentially activated pathways. Differentially activated pathways for each phase predicted by the different omics. "Unique" refers to the total number of pathways that are identified by one or more of the omics.

In the context of specific phases, it is observed that the RB phase exhibits a lower number of differentially active pathways. This finding is consistent with previous analyses suggesting that RB is the phase that differs the least from the other phases. However, the decreased number of DARs in RB may also be attributed to its specific functional profile. Enrichment-based methods have indicated that RB is characterized by processes related to cell division, DNA replication, mismatch repair, and the cell cycle. These pathways, which are associated to global cellular processes rather than specific reactions, are not included in MAMBA's metabolic model and therefore not captured in the analysis.

However, MAMBA does provide insights into downstream metabolic processes that may be associated with these general functions, such as folate biosynthesis, highlighted by both RNA-seq and ChIP-seq in RB. Folate biosynthesis plays a crucial role in various cellular processes related to DNA replication, including DNA synthesis, repair, and methylation reactions (Bailey & Gregory, 1999). OX phase also exhibits a similar pattern. General processes related to protein synthesis, such as RNA polymerase or ribosome biogenesis are not observed, but MAMBA still

captures the related underlying metabolic processes, including nucleotide, amino acid and tRNA biosynthesis.

Regarding RC (Fig. 28), the importance of energetic metabolism is maintained, as indicated by the high PES of glycolysis, fatty acid degradation, TCA cycle and pyruvate metabolism. As observed in the other phases, some global processes revealed by FE analysis, such as peroxisome, proteasome, autophagy or endocytosis are not present in MAMBA's prediction.

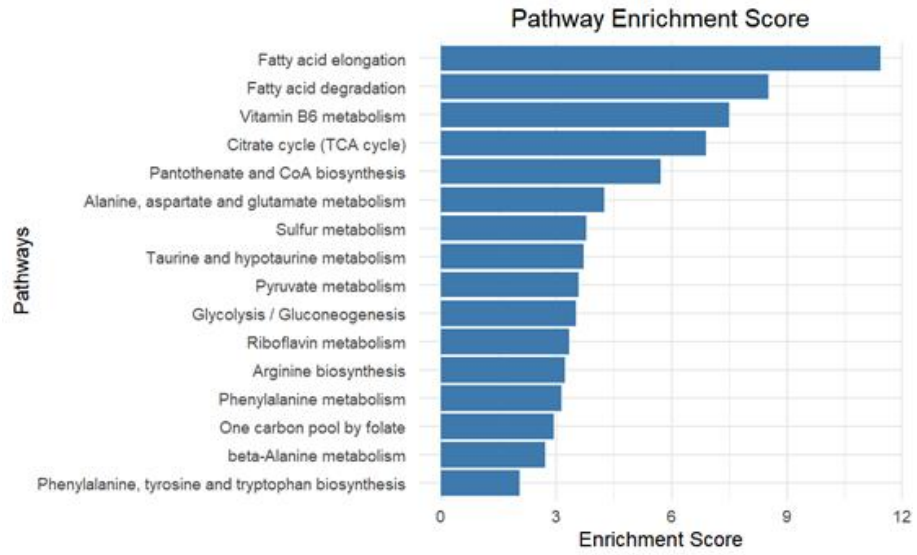


Figure 28: Pathway Enrichment Score. Percentage of reactions exclusively active in RC.

Despite the aforementioned limitation, MAMBA proved to be a valuable tool for identifying specific downstream processes that were not be observed through traditional FE analysis. In the case of the RC phase, MAMBA analysis revealed processes related to B-vitamin metabolism, including B2, B5, B9, and B6.

For instance, B5, also known as pantothenate, serves as a key precursor for the biosynthesis of coenzyme A (CoA). CoA plays a crucial role in various metabolic pathways, including the synthesis or degradation of fatty acids, phospholipids, and the operation of the TCA cycle (Leonardi & Jackowski, 2007). Vitamin B2, or riboflavin, is another Bvitamin that plays a vital role in cellular metabolism. It is essential for the formation of flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD), which act as coenzymes involved in oxidationreduction reactions in primary metabolic pathways, such as CTA cycle, oxidation, and degradation of amino acids (Lienhart *et al.*, 2013). The involvement of these compounds in such metabolic processes highlights its significance in the RB phase, strongly characterized by energy production. The identification of the onecarbon pool by folate as an important pathway in the RC phase, along with the observation of folate (vitamin B9) biosynthesis in the RB phase, suggests the significance of folate metabolism throughout the YMC. The one-carbon pool by folate pathway is responsible for the transfer of one-carbon units in the form of methyl groups (CH3) for various biochemical reactions, including nucleotide synthesis, amino acid metabolism, and the methylation of DNA, RNA, proteins, and other molecules. The methylation reactions facilitated by folate metabolism play a role in regulating gene expression and cellular signalling, which may be important for coordinating metabolic activities during this phase.

In summary, the analysis indicated that MAMBA provides complementary information to traditional enrichment methods for the functional profiling of the cycle. While enrichment methods capture differences in global processes such as protein synthesis, cell division or energy production, MAMBA reveals specific metabolic pathways that underlie these broader biological functions.

When comparing the prediction results based on different omics data, both transcriptomic

and epigenetic data demonstrate similar sensitivity and detect a similar number of pathways. Interestingly, when we examine the pathways identified by RNA-seq and ChIP-seq across all three phases (Fig. 29), we find that there are no pathways exclusively identified by RNA-seq. However, the overlap between the phases is only partial when considering them independently. This suggests that the observed differences between the two omics approaches are more likely due to biological delays between the different molecular layers, but a more in-depth analysis would be necessary to derive meaningful biological conclusions from these observations.

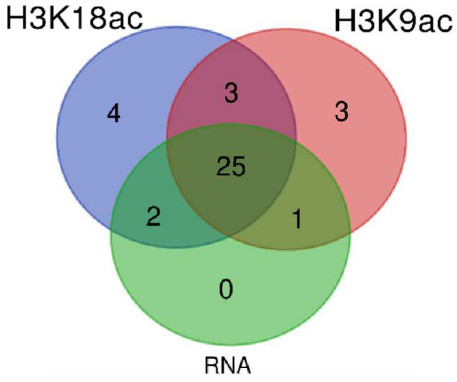


Figure 29: Overlap of relevant pathways. Overlap of the differentially activated pathway between the omics when considering all phases together.

4.6 Gene Set Variation Analysis

In order to evaluate the performance of both methods in terms of identifying and capturing the biological differences between conditions, a Gene Set Variation Analysis (GSVA) was applied only to those pathways identified as relevant by the different methods. GSVA provides an estimate of pathway activity by transforming gene-by-sample expression data into a corresponding gene-set-by-sample expression data, that can be used for clustering in a pathway-centric manner (Hänzelmann *et al.*, 2013).

For MAMBA, pathways with a PES higher than the significance threshold (general PES score) were considered relevant. In the case of FE methods, relevant pathways were defined as the pathways identified ORA with a False Discovery Rate lower than 0.05. The resulting GSVA scores, derived from the gene expression data of genes involved in pathways classified as relevant, were utilized for unsupervised clustering of the samples.

The unsupervised clustering analysis of the RNA-seq data (Fig. 30) showed distinct separation of samples according to the three phases of the cycle, indicating that the pathways identified as relevant by MAMBA and ORA were able capture the major biological changes across conditions. The only exception is Time 11, which is grouped within the RC phase in MAMBA based clustering.

In the case of ORA (Fig. 30), pathways can be clustered into 3 categories, each showing significantly higher activity in one of the phases compared to the other two. Aligning with previous work, the pathways enriched in the OX phase are related to protein synthesis, those enriched in the RB phase are associated with DNA replication, and the pathways enriched in the RC phase are involved in energetic metabolism. The clustering of the samples suggests a subdivision of RB in "early" RB (time points 6, 7 and 8), with a functional profile more similar to OX, and "late" RB (time points 9, 10 and 11), more similar to RC. Notably, in the case of MAMBA (Fig. 30), the pathways were clustered into two distinct groups instead of three. One group exhibited higher activity in the samples from RC, while the other group showed higher activity in the samples from OX. Interestingly, the pathways with higher activity in the RB

phase were shared between these two groups. This, on the one hand, indicates that RB is acting as an intermediate phase between OX and RC. However, it also highlights that the differential activity of RB, specifically associated with cell division, is not adequately captured by MAMBA.

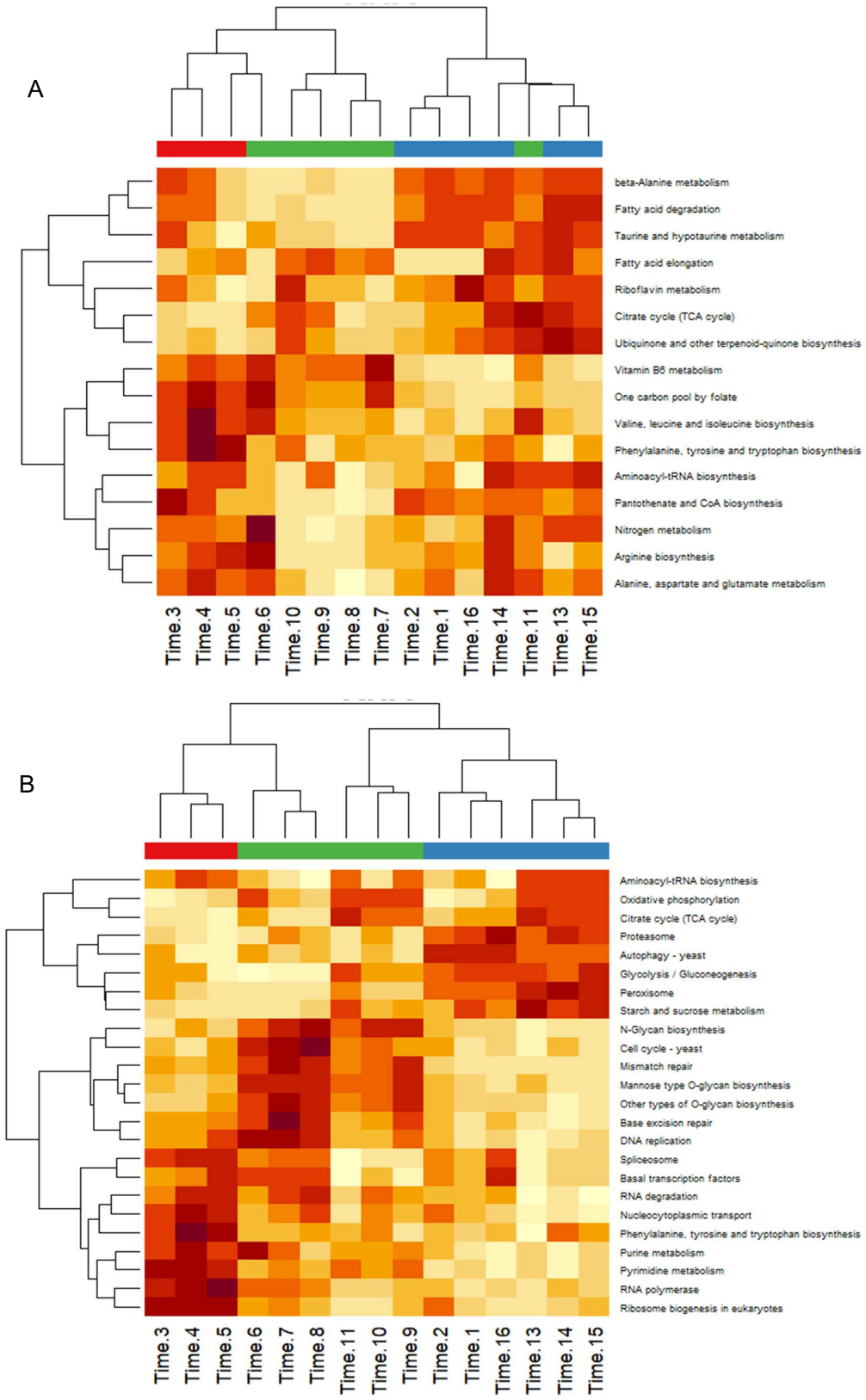


Figure 30: Unsupervised clustering of the samples using GSVA scores. Applied only to relevant pathways identified by MAMBA (A) and ORA (B) based on RNA-seq information.

In contrast to the clustering results based on ORA of ChIP-seq data, which did not reveal distinct phase divisions, the clustering based on pathways identified by MAMBA successfully grouped the samples according to their biological conditions. This indicates that MAMBA is better able to capture the relevant information from ChIP-seq (Fig. 31 and 32).

However, it is worth noting that there were a few exceptions in the clustering results. Specifically, Time 6, which falls on the border between OX and RB phases, was clustered within the OX phase, while Time 11, which falls on the border between RB and RC phases, was clustered within the RC phase. These observations indicate that the transitional samples at the phase boundaries may exhibit characteristics more similar to the adjacent phases and can be therefore problematic when treated as replicates.

In conclusion, the analysis of the transcriptomic data using both MAMBA and ORA demonstrated their ability to identify pathways that accurately captured the underlying biological differences between conditions. These findings were evident in the clear separation of samples into distinct phases, highlighting the robustness of both methods in capturing dynamic changes associated with the YMC. However, when it comes to the analysis of epigenetic data, a difference between MAMBA and ORA can be observed. MAMBA effectively identified pathways that exhibited differential activity across conditions, allowing for clustering of the sample following the 3 functional phases of the cycle, with only a few exceptions for samples at the border of RB. In contrast, the ORA of the epigenetic data failed to provide clear phase division, indicating its limitations in capturing the regulatory patterns associated with ChIP-seq data.

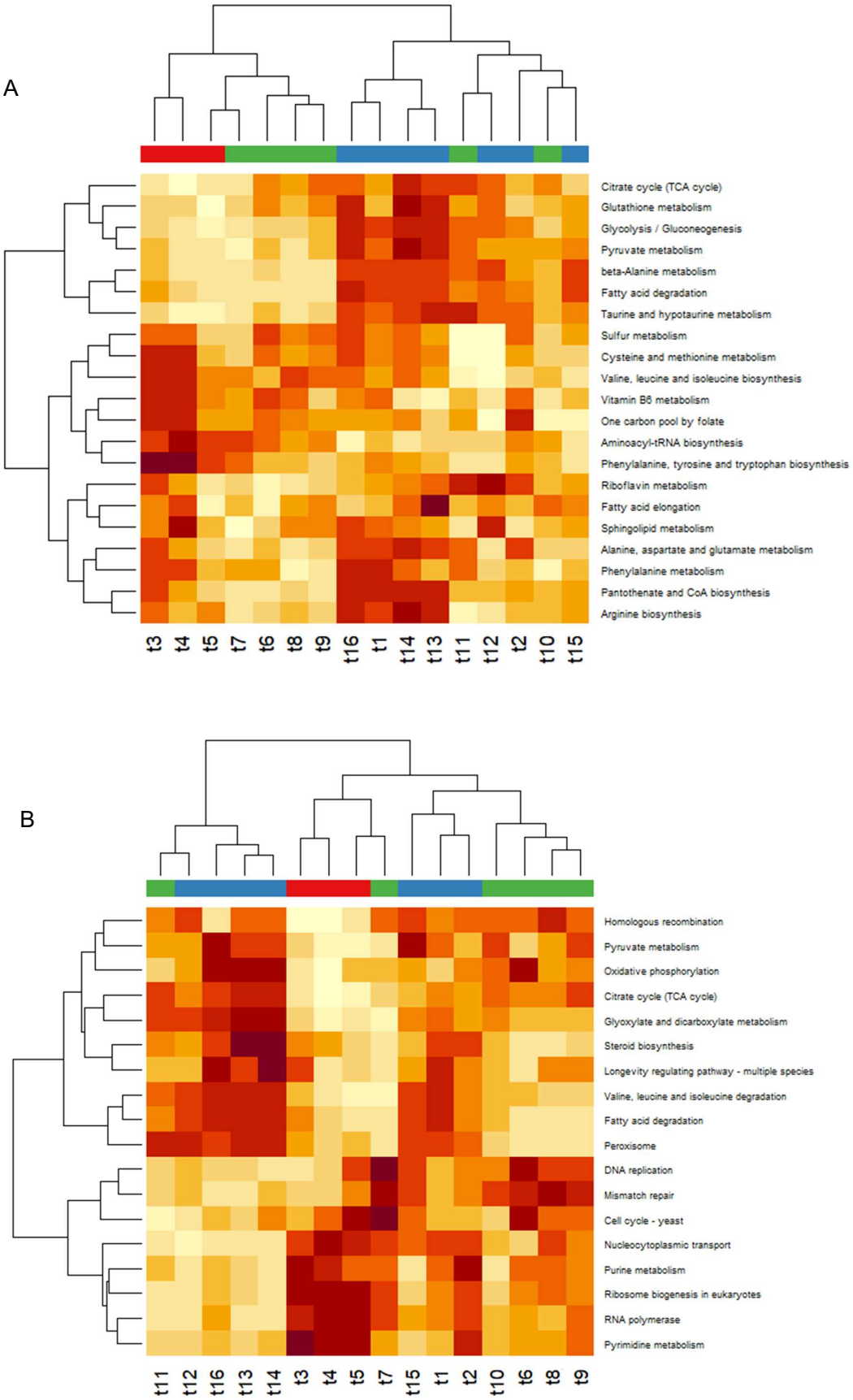


Figure 31: Unsupervised clustering of the samples using GSVA scores. Applied only to relevant pathways identified by MAMBA (A) and ORA (B) based on H3K18ac information.

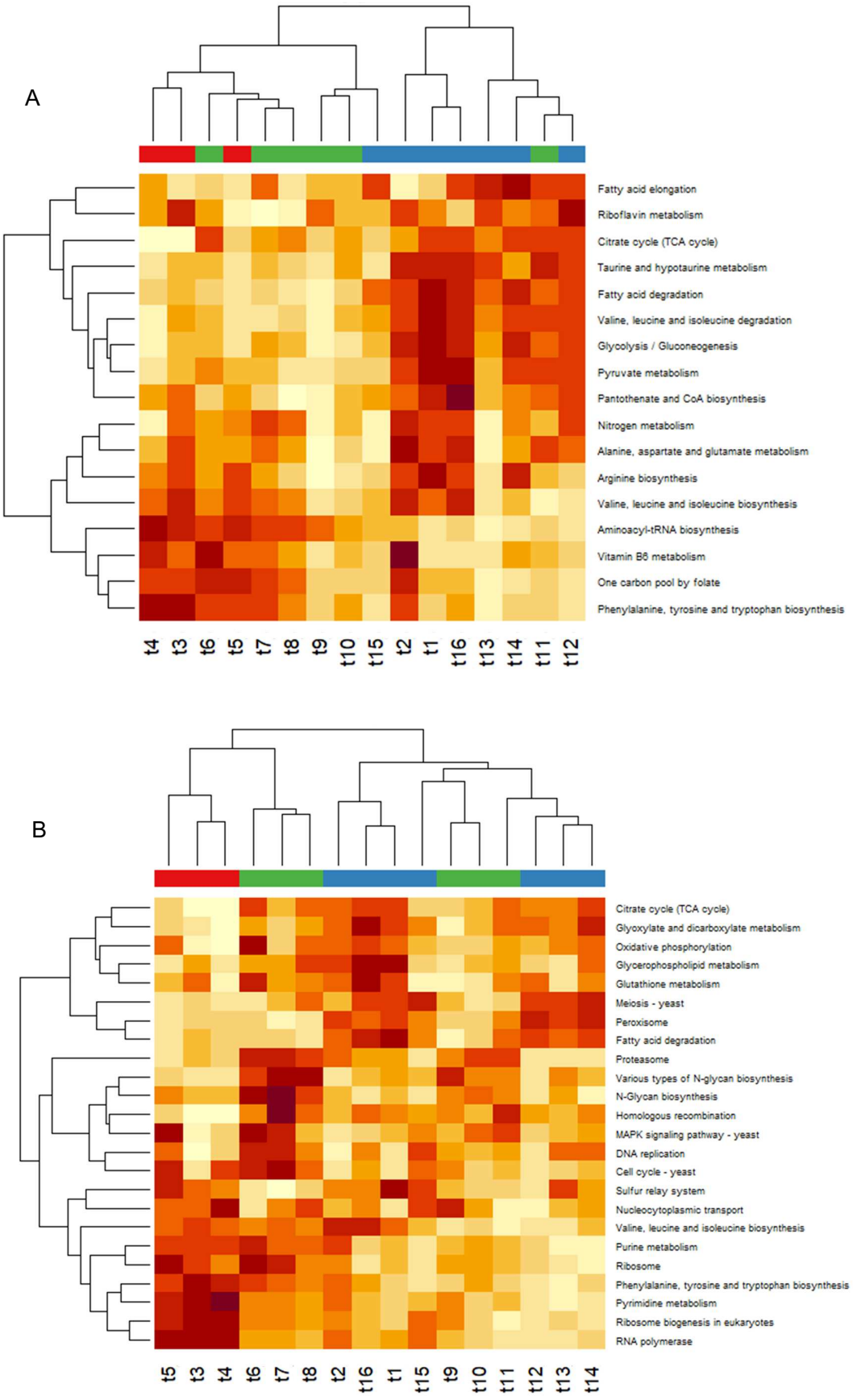


Figure 32: Unsupervised clustering of the samples using GSVA scores. Applied only to relevant pathways identified by MAMBA (A) and ORA (B) based on H3K9ac information.

5 Conclusions

In conclusion, this study conducted a comprehensive evaluation of MAMBA as a metabolic modelling tool, assessing its performance when applied to transcriptomic and epigenetic data. Furthermore, a comparison was made between MAMBA and conventional functional enrichment methods commonly used in functional analysis. Based on the findings, the following conclusions can be drawn:

1. A high-quality multi-omic dataset comprising transcriptomic, epigenomic, and metabolomic information of *Saccharomyces cerevisiae* during the Yeast Metabolic Cycle was successfully obtained. Differential expression analysis was effectively applied to the three transitions of the cycle.

2. Functional enrichment analysis enabled the identification of enriched biological functions associated with the different phases. The comparison between RNA-seq and ChIP-seq data revealed that RNA-seq demonstrated higher sensitivity in the context of Gene Set Enrichment Analysis, while Over Representation Analysis (ORA) yielded a similar number of relevant pathways across all three transitions. Moreover, the transcriptomic and epigenetic information showed distinct activation patterns in certain pathways, indicating a temporal shift between the two molecular layers.

3. MAMBA's metabolic modelling and flux predictions provided valuable insights into the functional profiling of the Yeast Metabolic Cycle, both for RNA-seq and ChIP-seq data. By translating quantitative gene-centric information into binary gene states, MAMBA compensated for the different dynamic ranges of the two technologies, yielding similar results in terms of sensitivity.

4. Overall, both MAMBA and conventional enrichment-based methods offered complementary information. ORA demonstrated the ability to identify broader biological functions, while MAMBA's results focused on the underlying metabolic pathways, as it is limited to the reactions that are explicitly included in the model. When utilizing RNA-seq data, both methods effectively captured the underlying biological differences between the phases. Notably, MAMBA exhibited superior performance when employing epigenetic information.

5.1 Limitations of the study

Several limitations were identified during the course of this research, which should be taken into consideration when interpreting the findings:

Firstly, MAMBA is characterized by a relative rather than absolute approach for metabolic modelling, as it relies on the differential values between conditions. The comparing conditions employed in this study are the phases of the YMC, which differ in very broad cellular functions, such as cell cycle or protein synthesis. MAMBA may not be the optimal tool for comparing such conditions, as the main expected differences are only indirectly included in the metabolic model.

Secondly, the absence of biological replicates within the same phase poses a limitation. Not all time points within a given phase, particularly those located at the phase boundaries, exhibited equivalent gene expression profiles. Incorporating biological replicates would have provided a more robust foundation for the analysis and enhanced the reliability of the results. Furthermore,

limitations were encountered during the functional analysis of MAMBA's results. A considerable amount of information was lost due to the inability to assign KEGG IDs to approximately half of the reactions encompassed by the metabolic model. This limitation restricts the biological interpretation of the results, as the comprehensive annotation of reactions would have facilitated a deeper understanding of the underlying metabolic pathways.

5.2 Future perspectives

MAMBA emerges as a promising approach for leveraging both transcriptomic and epigenetic data to unravel the complex dynamics of biological processes such as the Yeast Metabolic Cycle. However, there is ample room for further refinement and optimization, mainly addressing the following areas: Firstly, improving the functional annotation of the reactions encompassed by the metabolic model is crucial. Enhancing the coverage and accuracy of reaction annotations would significantly enhance the biological interpretation of the results, allowing for a more comprehensive understanding of the underlying metabolic pathways and their associated functions. Additionally, developing modifications to the incorporation of epigenetic data that account for the biological delay between chromatin modifications, transcriptional events, and metabolic regulation to provide a more realistic representation of the temporal dynamics of these processes and enhance the precision and reliability of the predictions generated by MAMBA.

6 References

Bailey, L. B., & Gregory, J. F. (1999). Folate Metabolism and Requirements1,2. *The Journal of Nutrition*, *129*(4), 779–782. Retrieved 2023-05-23, from https://www.sciencedirect.com/science/article/pii/S0022316623020126 doi: 10.1093/jn/129.4.779

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*, *42*(4), 426–437. Retrieved 2023-05-24, from https://www.sciencedirect.com/science/article/pii/S1097276511003327 doi: 10.1016/j.molcel.2011.05.004

Casaní Galdón, S. (2021). *Bioinformatic approaches to study the metabolic effect on Gene Regulation* (doctoral thesis). Retrieved 2023-05-15, from https://roderic.uv.es/handle/10550/77596 (Accepted: 2021-02-04T10:55:35Z)

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, *34*(17), i884–i890. doi: 10.1093/bioinformatics/bty560

Clish, C. B. (2015). Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harbor Molecular Case Studies*, *1*(1), a000588. Retrieved 2023-05-25, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4850886/ doi: 10.1101/mcs.a000588

Covert, M. W., & Palsson, B. O. (2003). Constraints-based models: Regulation of Gene Expression Reduces the Steady-state Solution Space. *Journal of Theoretical Biology*, *221*(3), 309–325. Retrieved 2023-06-01, from https://www.sciencedirect.com/science/article/pii/S0022519303930712 doi: 10.1006/jtbi.2003.3071

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. Retrieved 2023-05-11, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/ doi: 10.1093/bioinformatics/bts635

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., … Waldron, L. (2021). Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, *22*(1), 545–556. Retrieved 2023-05-10, from https://doi.org/10.1093/bib/bbz158 doi: 10.1093/bib/bbz158

Gowda, G. A. N., & Raftery, D. (2021). NMR Based Metabolomics. *Advances in experimental medicine and biology*, *1280*, 19–37. Retrieved 2023-05-25, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8816450/ doi: 10.1007/978-3-030-51652-9$_2$

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. Retrieved 2023-05-11, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615629/ doi: 10.1093/nar/gkn923

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, *14*(1), 7. Retrieved 2023-05-03, from https://doi.org/10.1186/1471-2105-14-7 doi: 10.1186/1471-2105-14-7

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and re-

cent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. Retrieved 2023-05-25, from https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (Publisher: Royal Society) doi: 10.1098/rsta.2015.0202

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLOS Computational Biology*, *11*(8), e1004321. Retrieved 2023-05-22, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004321 (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1004321

Kuang, Z., Cai, L., Zhang, X., Ji, H., Tu, B. P., & Boeke, J. D. (2014). High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nature Structural & Molecular Biology*, *21*(10), 854–863. doi: 10.1038/nsmb.2881

LaMar, D. (2015). FastQC.
doi: https://qubeshub.org/resources/fastqc

Leonardi, R., & Jackowski, S. (2007). Biosynthesis of Pantothenic Acid and Coenzyme A. *EcoSal Plus*, *2*(2). doi: 10.1128/ecosalplus.3.6.3.4

Lienhart, W.-D., Gudipati, V., & Macheroux, P. (2013). The human flavoproteome. *Archives of Biochemistry and Biophysics*, *535*(2), 150–162. Retrieved 2023-05-23, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3684772/ doi: 10.1016/j.abb.2013.02.015

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. Retrieved 2023-05-17, from https://doi.org/10.1186/s13059-014-0550-8 doi: 10.1186/s13059-014-0550-8

Maleki, F., Ovens, K., Hogan, D. J., & Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in Genetics*, *11*. Retrieved 2023-05-24, from https://www.frontiersin.org/articles/10.3389/fgene.2020.00654

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. Retrieved 2023-05-11, from https://journal.embnet.org/index.php/embnetjournal/article/view/200 (Number: 1) doi: 10.14806/ej.17.1.200

Marín de Mas, I. (2018). Chapter Sixteen - Multiomic Data Integration and Analysis via Model-Driven Approaches. In J. Jaumot, C. Bedia, & R. Tauler (Eds.), *Comprehensive Analytical Chemistry* (Vol. 82, pp. 447–476). Elsevier. Retrieved 2023-05-25, from https://www.sciencedirect.com/science/article/pii/S0166526X18300692 doi: 10.1016/bs.coac.2018.07.005

Mehrmohamadi, M., Sepehri, M. H., Nazer, N., & Norouzi, M. R. (2021). A Comparative Overview of Epigenomic Profiling Methods. *Frontiers in Cell and Developmental Biology*, *9*. Retrieved 2023-05-10, from https://www.frontiersin.org/articles/10.3389/fcell.2021.714687

Mellor, J. (2016). The molecular basis of metabolic cycles and their relationship to circadian rhythms. *Nature Structural & Molecular Biology*, *23*(12), 1035–1044. Retrieved 2023-05-24, from https://www.nature.com/articles/nsmb.3311 (Number: 12 Publisher: Nature Publishing Group) doi: 10.1038/nsmb.3311

Mosegaard, S., Dipace, G., Bross, P., Carlsen, J., Gregersen, N., & Olsen, R. K. J. (2020). Riboflavin Deficiency—Implications for General Human Health and Inborn Errors of Metabolism. *International Journal of Molecular Sciences*, *21*(11), 3847. Retrieved 2023-05-23, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7312377/ doi: 10.3390/ijms21113847

Muhammad, I. I., Kong, S. L., Akmar Abdullah, S. N., & Munusamy, U. (2019). RNA-seq and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional

Regulatory Mechanism. *International Journal of Molecular Sciences*, *21*(1), 167. Retrieved 2023-05-10, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6981605/ doi: 10.3390/ijms21010167

Ng, R. H., Lee, J. W., Baloni, P., Diener, C., Heath, J. R., & Su, Y. (2022). Constraint-Based Reconstruction and Analyses of Metabolic Models: Open-Source Python Tools and Applications to Cancer. *Frontiers in Oncology*, *12*. Retrieved 2023-05-03, from https://www.frontiersin.org/articles/10.3389/fonc.2022.914594

Orth, J. D., Thiele, I., & Palsson, B. . (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. Retrieved 2023-05-24, from https://www.nature.com/articles/nbt.1614 (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/nbt.1614

O'Geen, H., Echipare, L., & Farnham, P. J. (2011). Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. In T. O. Tollefsbol (Ed.), *Epigenetics Protocols* (pp. 265–286). Totowa, NJ: Humana Press. Retrieved 2023-05-25, from https://doi.org/10.1007/978-1-61779-316-5$_2$0 doi: 10.1007/978-1-61779-316-5$_2$0

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. doi: 10.1093/bioinformatics/btq033

Rao, A. R., & Pellegrini, M. (2011). Regulation of the yeast metabolic cycle by transcription factors with periodic activities. *BMC Systems Biology*, *5*(1), 160. Retrieved 2023-05-25, from https://doi.org/10.1186/1752-0509-5-160 doi: 10.1186/1752-0509-5-160

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. doi: 10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. doi: 10.1093/bioinformatics/btp616

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. Retrieved 2023-05-08, from https://www.pnas.org/doi/10.1073/pnas.0506580102 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0506580102

Ugidos, M., Nuño-Cabanes, C., Tarazona, S., Ferrer, A., Nielsen, L. K., Rodríguez-Navarro, S., … Conesa, A. (2022). *MAMBA: a model-driven, constraint-based multiomic integration method.* bioRxiv. Retrieved 2023-05-03, from https://www.biorxiv.org/content/10.1101/2022.10.09.511458v1 (Pages: 2022.10.09.511458 Section: New Results) doi: 10.1101/2022.10.09.511458

Ugidos Guerrero, M. (2023). *Statistical Methods Development for the Multiomic Systems Biology* (Tesis doctoral, Universitat Politècnica de València). (Accepted: 2023-05-02T07:50:30Z) doi: 10.4995/Thesis/10251/193031

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, *10*(1), 57–63. Retrieved 2023-05-25, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/ doi: 10.1038/nrg2484

Zyla, J., Marczyk, M., Weiner, J., & Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics*, *18*(1), 256. Retrieved 2023-05-08, from https://doi.org/10.1186/s12859-017-1674-0 doi: 10.1186/s12859-017-1674-0

Supplementary material

Supplementary figures

PCA of ChIP-seq data before normalization

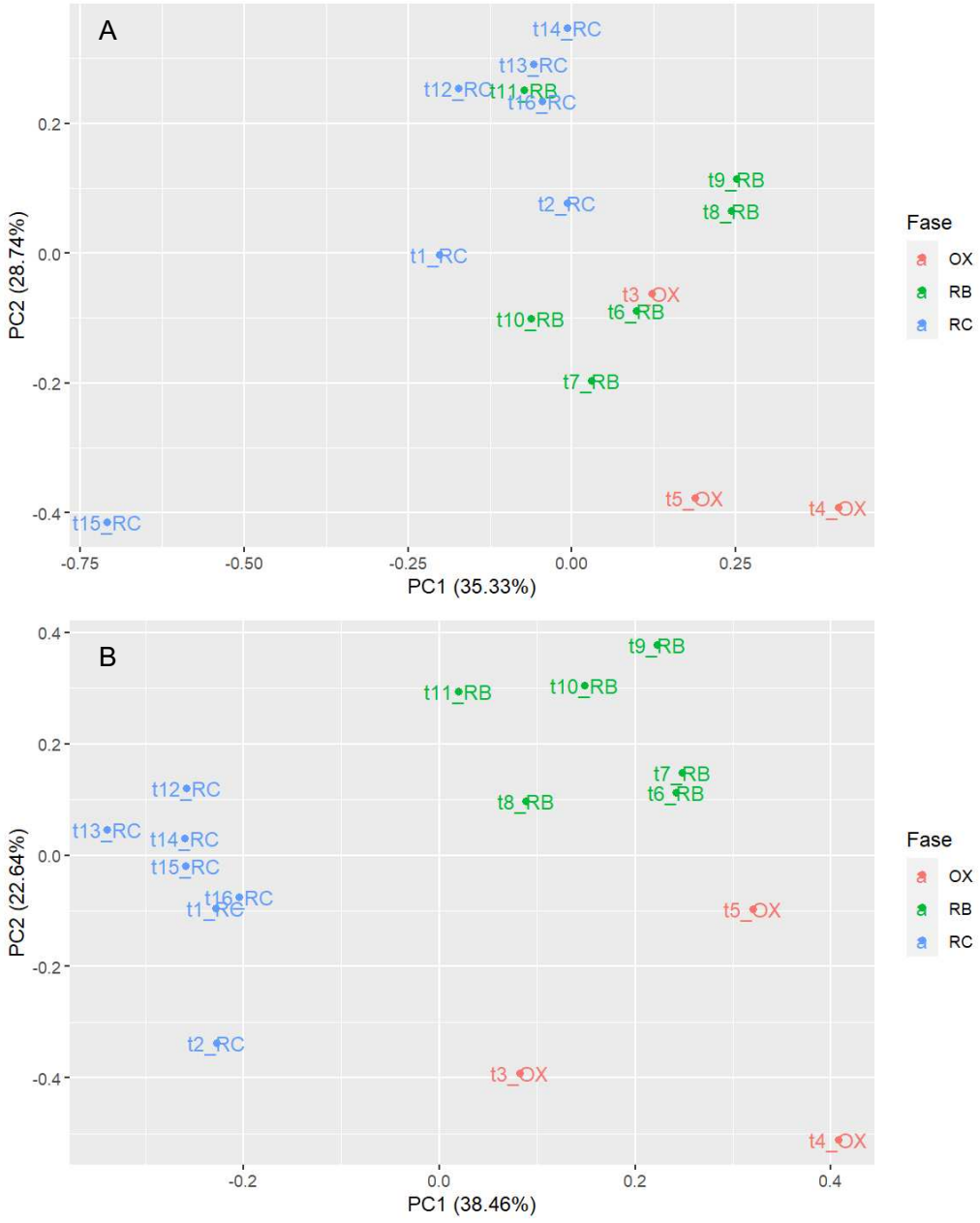


Figure 33: PCA of samples before H3 control normalization. Samples are coloured according to the corresponding phase of the cycle. Based on H3K18ac (A) and H3K9ac (B) ChIP-seq data.

Pathway Enrichment Score for MAMBA

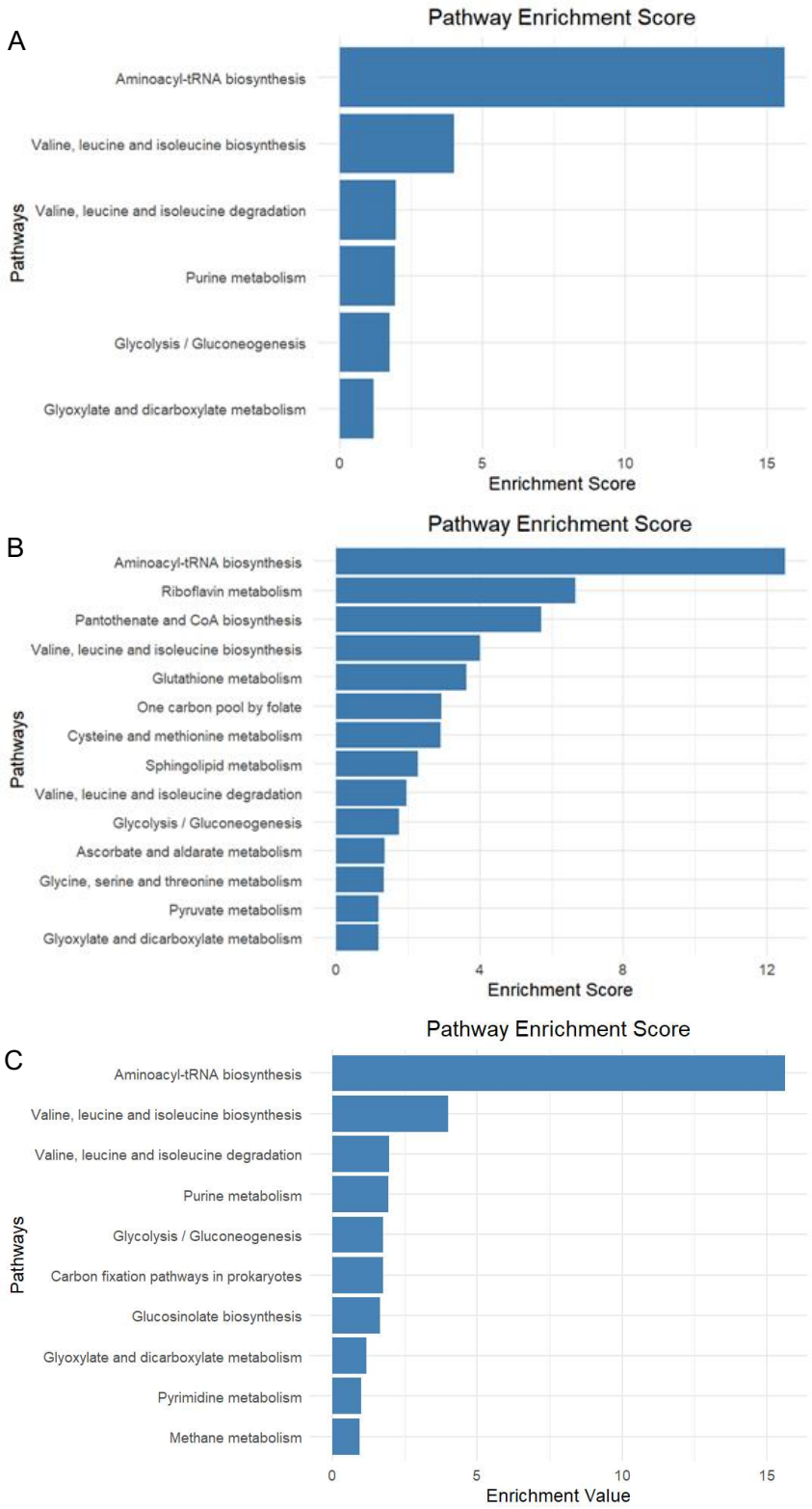


Figure 34: Pathway Enrichment Score (PES) in the OX phase. Relevant Pathways identified by MAMBA with the corresponding PES based on RNA (A), H3K18ac (B) and H3K18ac (C) information.

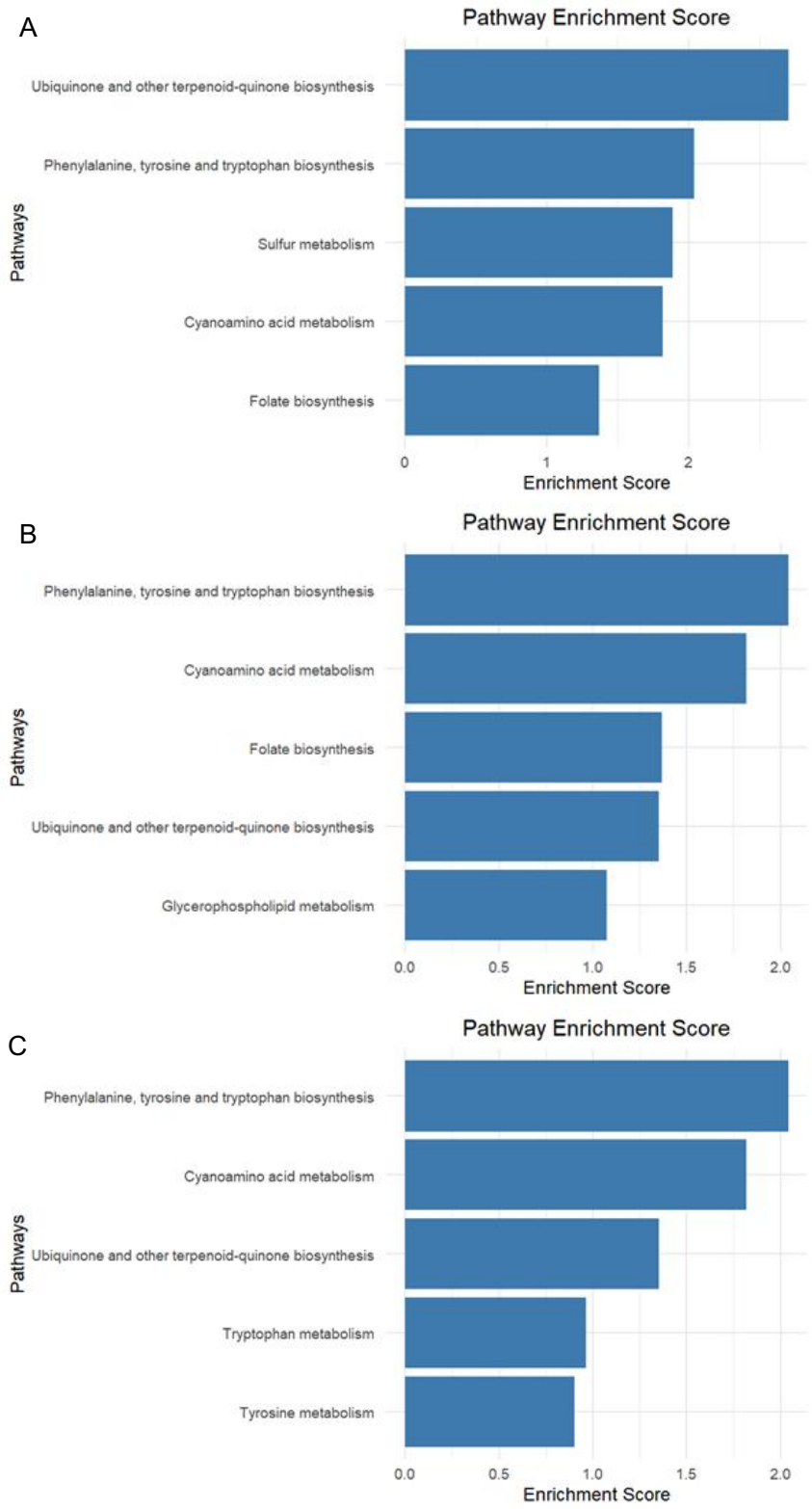


Figure 35: Pathway Enrichment Score (PES) in the RB phase. Relevant Pathways identified by MAMBA with the corresponding PES based on RNA (A), H3K18ac (B) and H3K18ac (C) information.

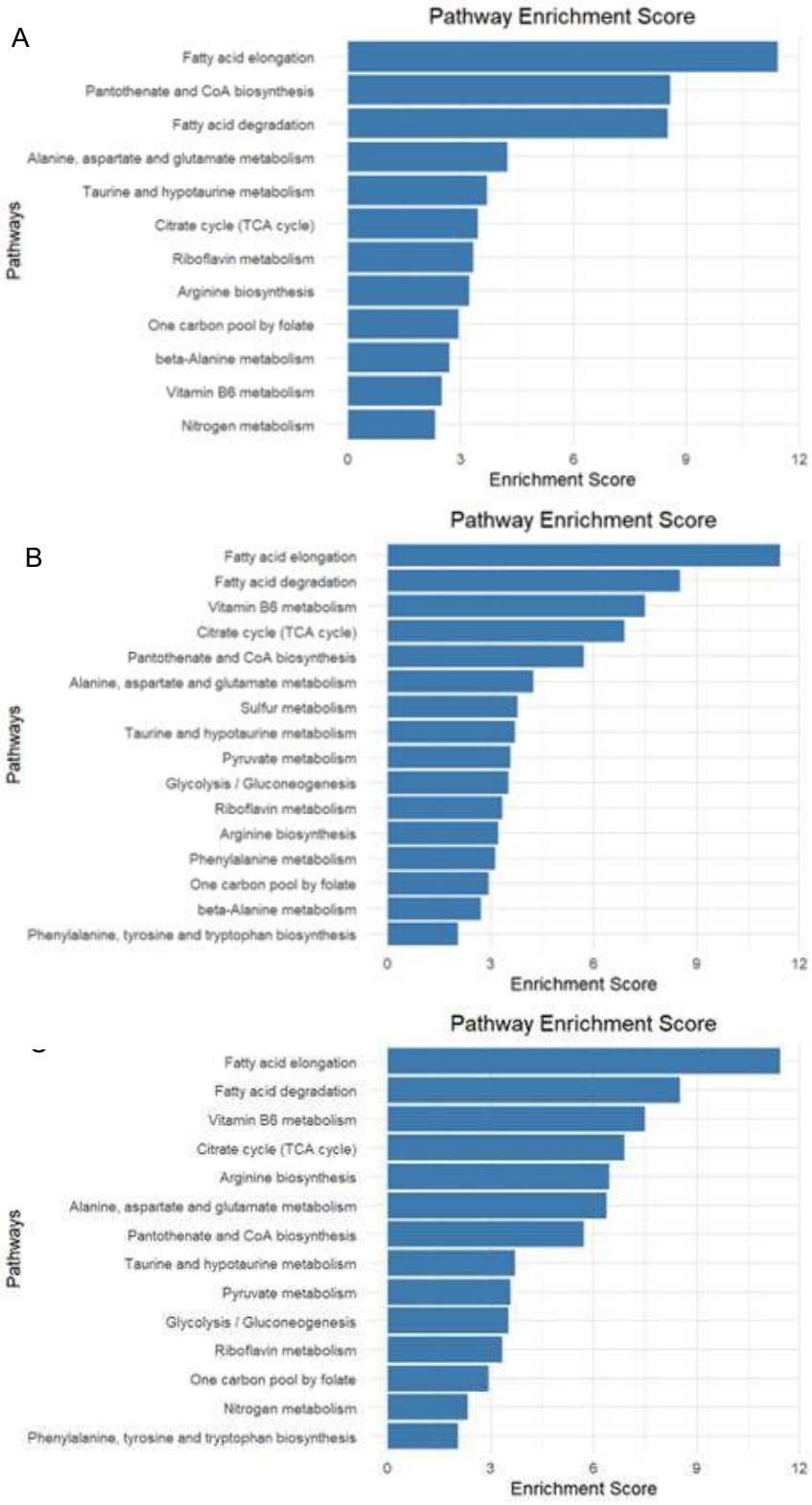


Figure 36: Pathway Enrichment Score (PES) in the RC phase. Relevant Pathways identified by MAMBA with the corresponding PES based on RNA (A), H3K18ac (B) and H3K18ac (C) information.

Sustainable Development Goals

This project is related to several Sustainable Development Goals (SDG) (Fig. 37):

- **Health and Well-being (SDG 3):** Developing strategies that enable to model cell metabolism based on experimental data is key for understanding the molecular mechanisms of complex diseases and creating effective treatments.
- **Climate Action (SDG 11)** and **Responsible production and consumption (SDG 12)**: In this work, experimental data were retrieved from public repositories. Analysing pre-existing data from a different scientific perspective allowed to draw significant biological conclusions while minimizing the production of plastic waste or other contaminants. In this sense, Bioinformatic approaches favour a more sustainable scientific when compared to wetlab.
- **Quality Education (SDG 4)** and **Gender Equality (SDG 5):** This work contributes the development and growth of girls and woman in science.

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Proceede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**			X	
ODS 5. **Igualdad de género.**			X	
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**				X
ODS 10. **Reducción de las desigualdades.**				X
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**		X		
ODS 13. **Acción por el clima.**		X		
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

Figure 37: Sustainable Development Goals (SDG). Each SDG has been assigned to a certain level of relationship with the project (High, Medium, Low, Not applicable).

R script

The R script used for functional analysis of MAMBA's flux prediction to obtain the biologically relevant pathways in the different phases is presented as an example of the bioinformatic workflow that has been followed in this work.

R script for functional analysis of MAMBA

2023-06-03

```
library(ggplot2)
library(KEGGREST)
library(GSVA)
library(RColorBrewer)
```

Para el análisis de flujos de MAMBA (H3K9ac)

PASOS PREVIOS GENERALES - Lo primero es sacar los ID de los pathways:

```
name_KEGG<-read.table("C:/Users/crist/Downloads/id_kegg.tsv",sep = "\t")
#Con esto tengo las reacciones, pero necesito los pathways

reaction<-name_KEGG$V2
pathway_id<-c()
no_annotated<-c()
for (i in reaction){
  result <- tryCatch(keggGet(i), error = function(x) {return(NA)})
  if (!is.na(result)){
    result_name<-names(result[[1]]$PATHWAY)
    if(is.null(result_name)){
      no_annotated<-append(no_annotated,i)
      }
    pathway_id<-append(pathway_id,result_name)
  }
}

KEGG_pathway_ID<-unique(pathway_id)
```

KEGG_pathway_ID –> todos los pathways con su KEGG ID no_annotated –> reacciones CON KEGG ID que no tienen información de a qué pathway pertenecen

- Una vez tenemos una lista con todos los ID de los PATHWAYS, queremos una lista con todos los ID de las reacciones que hay en ese pathway dentro.

```
pathway_reaction<-c()
no_reaction<-c()
for (m in 1:length(KEGG_pathway_ID)){
  z<-KEGG_pathway_ID[m]
  to<-keggGet(z)
  me<-names(to[[1]]$REACTION)
  if(is.null(me)){
      no_reaction<-append(no_reaction,m)
      }
  pathway_reaction[[KEGG_pathway_ID[m]]]<- me
}
```

AHORA YA ESPECÍFICO DE LA K9

1. Cargar los flujos y quedarte solo con los exclusivos en alguna condición (DAR)

```
fluxes_alejandro_k9<-read.csv("C:\\Users\\crist\\Downloads\\reaction_activity_k9.csv")

fluxes_alejandro_k9$total<-apply(fluxes_alejandro_k9[2:4],1,sum)
fluxes_dif<-fluxes_alejandro_k9[which(fluxes_alejandro_k9$total != 3 & fluxes_alejandro_k9$total != 0),]

fluxes_dif<-fluxes_alejandro_k9[which(fluxes_alejandro_k9$total ==1),]
```

2. Definir las DAR para las 3 condiciones. El orden es RC, OX, RB.

```
reactions_ox<-fluxes_dif[which(fluxes_dif$Condition_2 == 1),]
dif_reactions_ox<-as.vector(reactions_ox$Reaction_id)

reactions_rc<-fluxes_dif[which(fluxes_dif$Condition_1==1),]
dif_reactions_rc<-as.vector(reactions_rc$Reaction_id)

reactions_rb<-fluxes_dif[which(fluxes_dif$Condition_3 == 1),]
dif_reactions_rb<-as.vector(reactions_rb$Reaction_id)
```

3. Anotar las reacciones del modelo con el KEGG_ID correspondiente

```
name_KEGG<-read.table("C:/Users/crist/Downloads/id_kegg.tsv",sep = "\t")
all_dif_reactions<-list(RB=dif_reactions_rb,RC=dif_reactions_rc,
                        OX=dif_reactions_ox)

KEGG_ID<-c()
KEGG_ID_all<-list()
order<-c("RB","RC","OX")
for (i in order){
  for (s in all_dif_reactions[[i]]){
    index<-which(name_KEGG$V1 == s)
    id<-name_KEGG[index,2]
    KEGG_ID<-append(KEGG_ID,id)
  }
  KEGG_ID_all[[i]]<-KEGG_ID
  KEGG_ID<-c()
}
```

4. Calcular el Pathway Enrichment Score: previamente se ha generado una lista con todos los pathways y todas las reacciones que hay dentro (pathway_reaction)

```
PSE_list_all<-list()

PES_new<-data.frame()
PES_all<-list()
for (i in order){
  for (r in 1:length(pathway_reaction)){
      e<-intersect(pathway_reaction[[r]],KEGG_ID_all[[i]])
      v<-length(pathway_reaction[[r]])
      y<-length(e)
      PES_score<-y/v*100
      PES_df<-data.frame("pathway"= names(pathway_reaction)[r], "PES" = PES_score, "phase" = i)
      PES_new<-rbind(PES_new, PES_df)
  }
  PES_all[[i]]<-PES_new
  PES_new<-data.frame()
```

```
}
```

3. Quedarnos solo con aquellos pathways que tengan un PES significativo

```
for (i in order){
  PES_all[[i]]<-PES_all[[i]][which(PES_all[[i]]$PES !=0),]
}
```

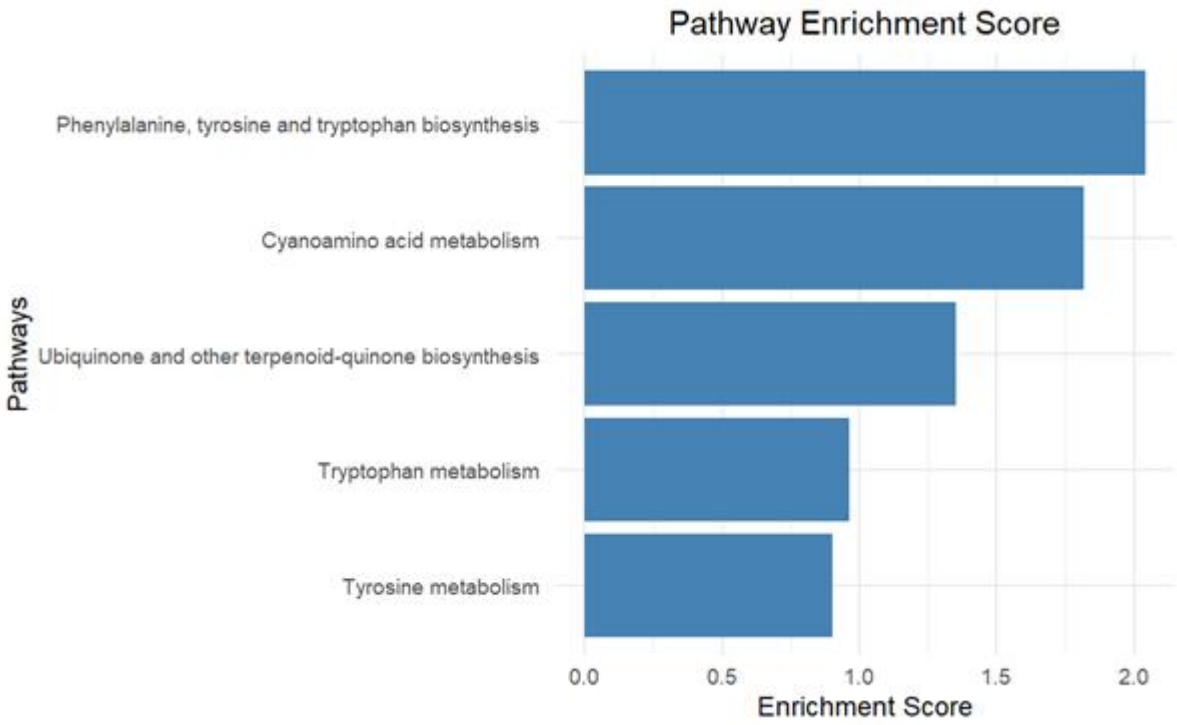
4. Extraer el nombre biológico descriptivo descriptivo de cada identificador. -cambiar el identificador de rn (general) a sce (de Saccharomyces Cerevisiae)
   - consulta de keggGet para saber el "name"
   - crear una data frame con todos los rn, sce y biological pathway
   - quedarse solo con los que tengan las tres cosas

```
df_ID_all<-data.frame()
for (i in order){
  for (p in names(pathway_reaction)){
    sce<-gsub("rn","sce",p)
    kegg_result<-tryCatch(keggGet(sce),error=function(x){return(NA)})
    if(!is.na(kegg_result))
      name<-as.character(kegg_result[[1]][["NAME"]])
      name_ok<-strsplit(name, "- Sac")[[1]][1]
      df_ID<-data.frame("sce"=sce,"biological_name"=name_ok,"pathway"=p)
      df_ID_all<-rbind(df_ID_all, df_ID)
      name<-"Hola"
    }
  df_ID_all<-unique(df_ID_all)
}
```

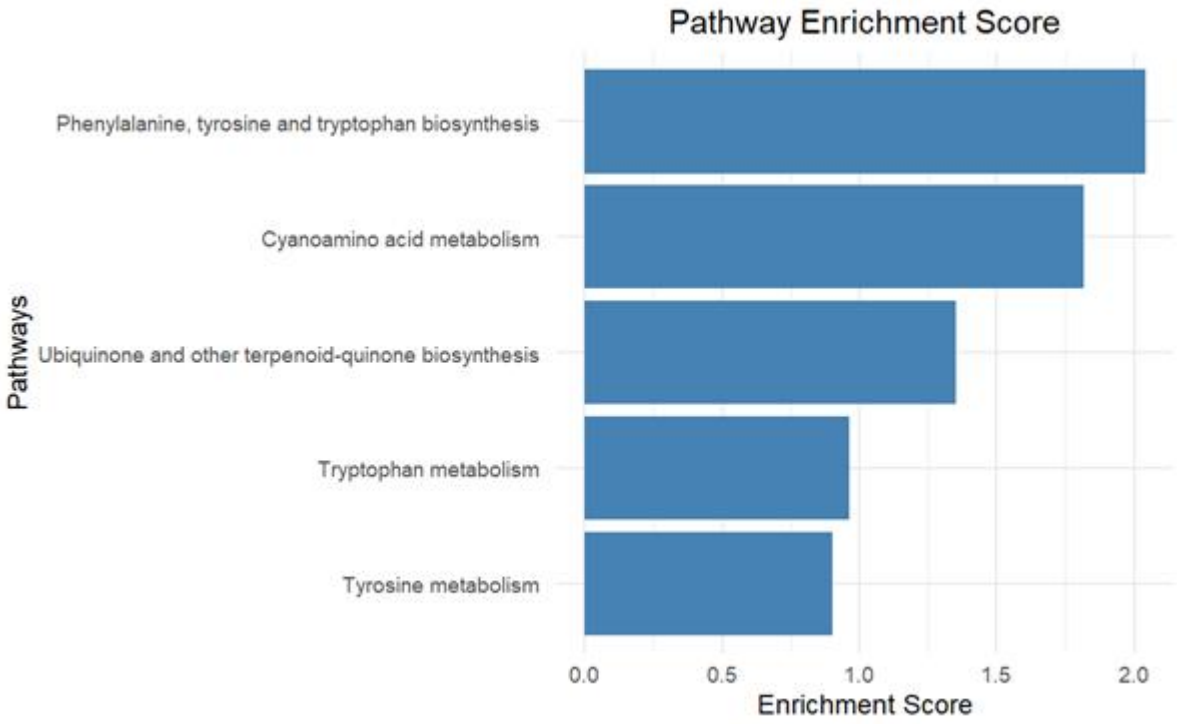
```
df_merged_all<-list()
final_df_all<-list()
for (i in order){
  df_merged_all[[i]]<-merge(PES_all[[i]],df_ID_all,by="pathway", all=TRUE)
  final_df_all[[i]]<-unique(na.omit(df_merged_all[[i]][which(df_merged_all[[i]]$phase ==

 i ),]))
}

for (i in order){
  final_df_all[[i]]<-final_df_all[[i]][which(final_df_all[[i]]$biological_name != "Hola"),]
}
```

5. Hacer un plot para cada fase por separado:

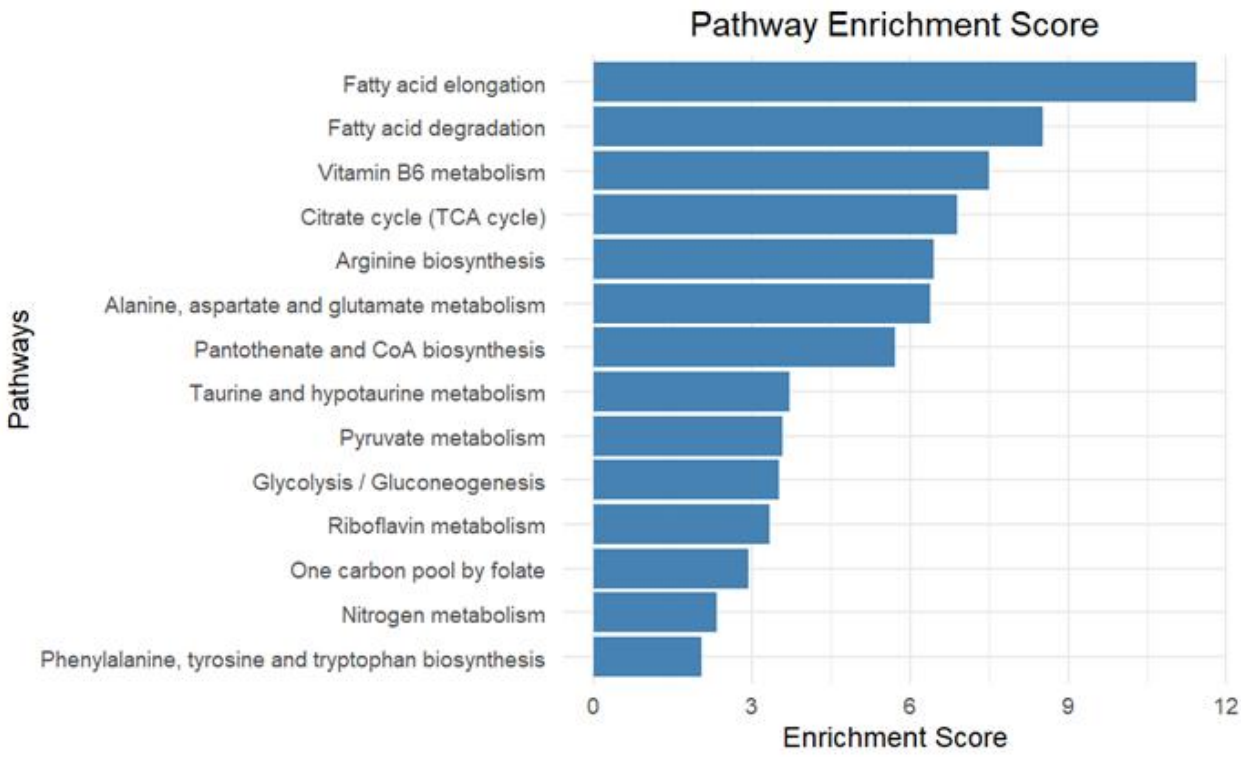
```
  ggplot(data = final_df_all[["OX"]], aes(x = PES, y = reorder(biological_name, PES))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Enrichment Score", y = "Pathways") +
  ggtitle("Pathway Enrichment Score") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(data = final_df_all[["RB"]], aes(x = PES, y = reorder(biological_name, PES))) +
geom_bar(stat = "identity", fill = "steelblue") +
labs(x = "Enrichment Score", y = "Pathways") +
ggtitle("Pathway Enrichment Score") +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5))
```



```
  ggplot(data = final_df_all[["RC"]][which(final_df_all[["RC"]]$PES>2),],

  aes(x = PES, y = reorder(biological_name, PES))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Enrichment Score", y = "Pathways") +
  ggtitle("Pathway Enrichment Score") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



6. Quedarse solo con los pathways que tengan un PES mayor que el threshold

```
sig_pathways_sce_ox_k9<-final_df_all[["OX"]][which(final_df_all[["OX"]]$PES>2),]$sce
sig_pathways_sce_rb_k9<-final_df_all[["RB"]][which(final_df_all[["RB"]]$PES>2),]$sce
sig_pathways_sce_rc_k9<-final_df_all[["RC"]][which(final_df_all[["RC"]]$PES>2),]$sce

sig_pathways_sce_all_k9<-c(sig_pathways_sce_ox_k9,sig_pathways_sce_rb_k9,sig_pathways_sce_rc_k9)
```

7. Para el GSVA:

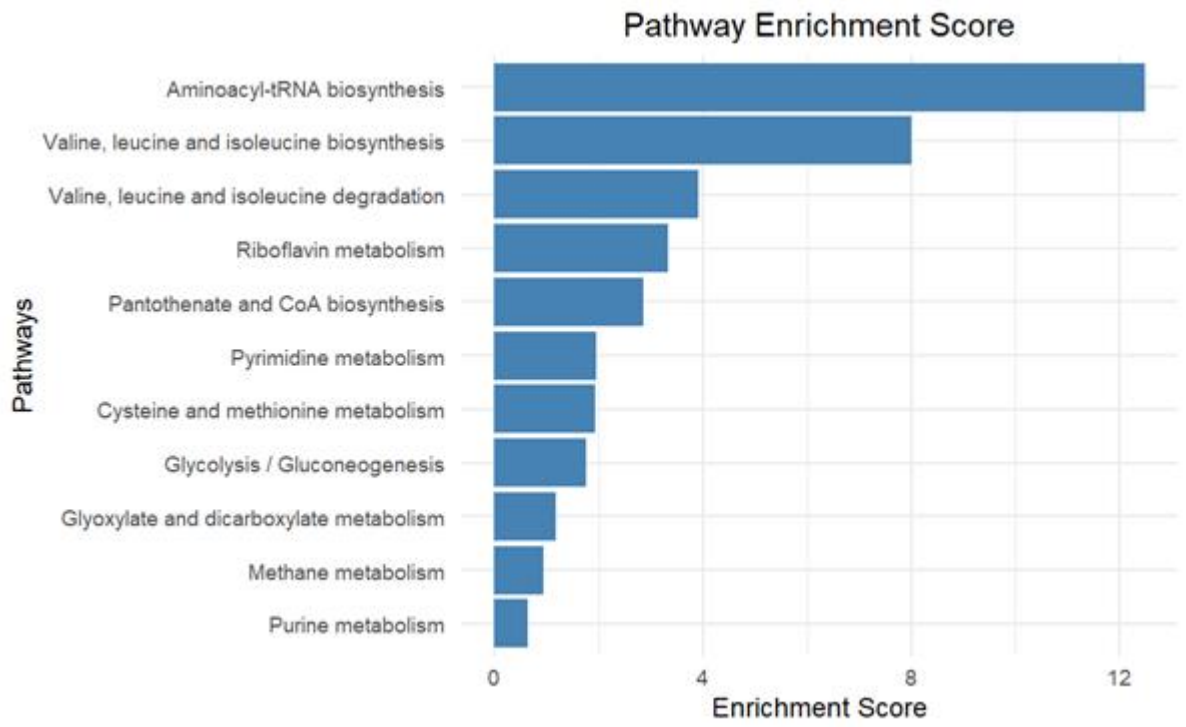
- sacar todos los genes que que esten dentro de los pathways significativos

```
sig_genes_pathways_k9<-list()
for (m in sig_pathways_sce_all_k9){
  bu<-tryCatch(keggGet(m),error=function(x){return(NA)})
    if(!is.na(bu)){
      ho<-bu[[1]][["GENE"]]
      s<-length(ho)
      l<-seq(from=1,to=s,by=2)
      d<-bu[[1]][["NAME"]]
      e<-strsplit(d, "- Sac")
      e<-e[[1]][1]
```

```
      sig_genes_pathways_k9[[e]]<-ho[1]
      e<-"Hola"
  }
}
```

- hacer un GSVA solo considerando los pathways significativos

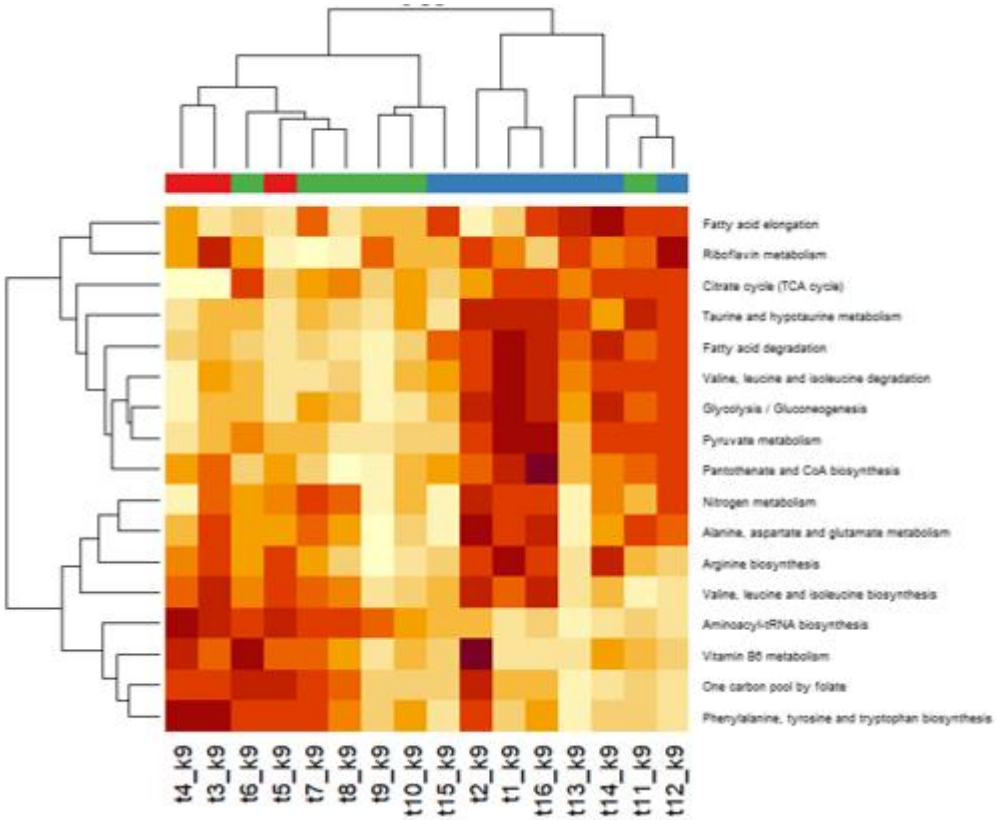
```
load("C:/Users/crist/Documents/crist_TFG/chip/countsK9_H3.RData")
boxplot(log(countsK9_H3))
```




```
gs<-sig_genes_pathways_k9
X<-countsK9_H3
gsva.es.k9 <- gsva(X, gs, verbose=TRUE)
```

- Heatmap (unsupervised clustering)

```
my_group<-c(1,1,1,3,3,3,3,3,3,2,2,2,2,2,2,2)
#my_group<-c(1,1,3,3,3,3,3,3,3,2,2,1,2,2,2,2,2)
colSide <- brewer.pal(9, "Set1")[my_group]
group_order<-c(10,1,2,3,4,5,6,7,8,9,11,12,13,14,15,16)

#group_order<-c(2,1,3,4,5,6,7,8,15,11,10,16,13,14,9,12)

heatmap(gsva.es.k9, ColSideColors=colSide,main= "K9", Colv = group_order,cexRow = 0.5)
```


Fatty acid elongation
Riboflavin metabolism
Citrate cycle (TCA cycle)
Taurine and hypotaurine metabolism
Fatty acid degradation
Valine, leucine and isoleucine degradation
Glycolysis / Gluconeogenesis
Pyruvate metabolism
Pantothenate and CoA biosynthesis
Nitrogen metabolism
Alanine, aspartate and glutamate metabolism
Arginine biosynthesis
Valine, leucine and isoleucine biosynthesis
Aminoacyl-tRNA biosynthesis
Vitamin B6 metabolism
One carbon pool by folate
Phenylalanine, tyrosine and tryptophan biosynthesis
t4_k9
t3_k9
t6_k9
t5_k9
t7_k9
t8_k9
t9_k9
t10_k9
t15_k9
t2_k9
t1_k9
t16_k9
t13_k9
t14_k9
t11_k9
t12_k9