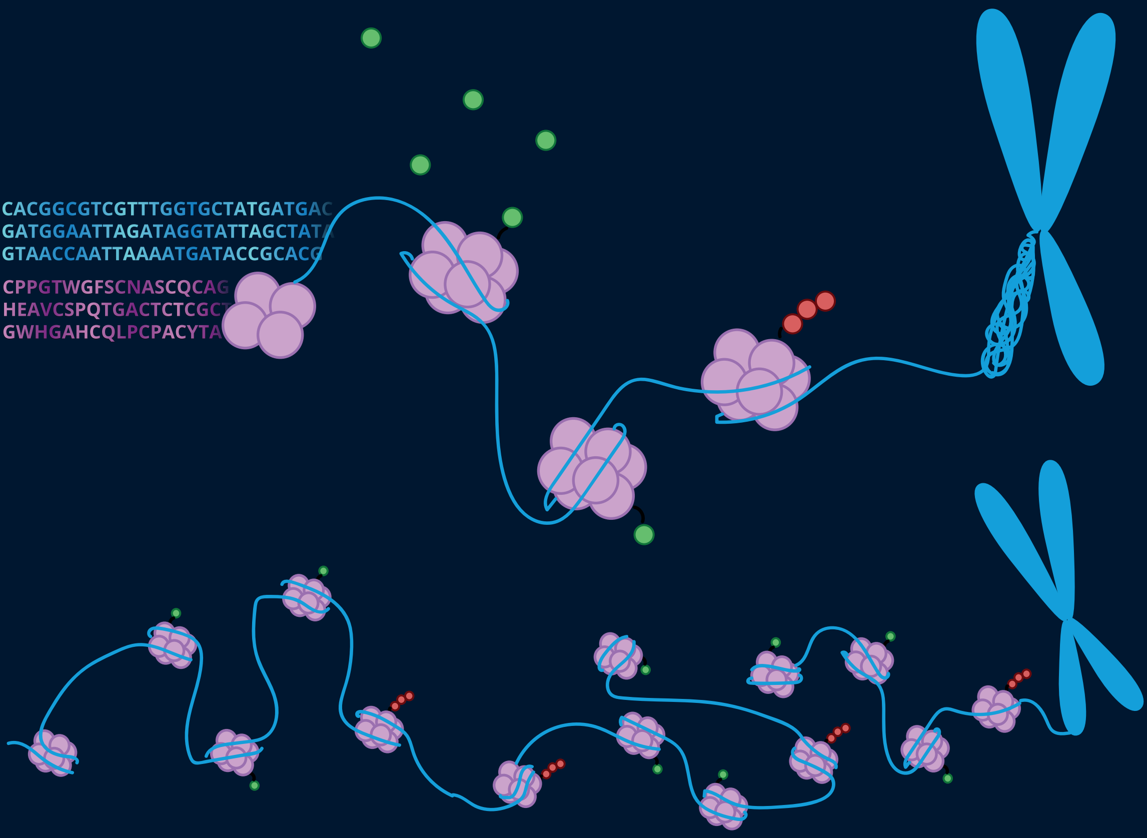
Bioinformatic Approaches to Study the Metabolic Effects on Gene Regulation

Salvador Casaní Galdón
PhD Thesis



CACGGCGTCGTTTGGTGCTATGATGAC
GATGGAATTAGATAGGTATTAGCTATA
GTAACCAATTAAAATGATACCGCACG
CPPGTWGFSCNASCQCAG
HEAVCSPQTGACTCTCGCT
GWHGAHCQLPCPACYTA

Vniversitat de València

Department of Information Technologies, Communications and Computing

Supervisor:
Dr. Ana Conesa Cegarra

December, 2020

Bioinformatic Approaches to Study the Metabolic Effects on Gene Regulation

Salvador Casaní Galdón

Supervisor:
Dr. Ana Conesa Cegarra
Tutor:
Vicente Arnau Llombart

A doctoral thesis submitted to

Department of Information Technologies, Communications and Computing

December 2020

Abstract

Cellular adaptation to changing environments constitutes a critical mechanism for cell survival. Cells primarily respond to external conditions by modulating the molecular mechanisms that regulate gene expression or protein activity, granting a rapid response to external metabolic changes. Therefore, metabolic sensing constitutes an important step in cell adaptation, and epigenetics is now considered the mechanism that connects metabolic shifts with gene regulation. Epigenetic marks give cells the capability of shaping chromatin conformation, which in turn regulates gene expression. Consequently, the correct functioning of a cell's epigenetic program is critical for cellular adaptation to changing conditions.

Different epigenetic modifiers rely on metabolite availability to modify the cell's epigenetic landscape. Recent studies point towards the accumulation of key metabolites as the critical mechanism by which epigenetic modifiers modulate the chromatin marks. This can be appreciated in circadian rhythms, where epigenetic changes mediate the cross-talk between metabolic oscillations and gene expression. Deficiencies that disconnect this molecular regulation lead to diseases, such as metabolic syndrome. The study of the metabolic control of the epigenome and transcriptome is an emerging field of research. Multiple studies have generated large, high-throughput datasets that measure gene expression, metabolites and histone modifications, among others, to study these interconnections; although a wealth of literature is accumulating, the precise mechanisms of these multi-layered regulations are still to be fully elucidated. Also, a consensus pathway describing these processes cannot yet be found in any of the common biological pathway databases. One critical need in the field is the integrative analysis of existing molecular data to propose detailed regulatory models for the interplay between metabolism, chromatin state and transcription.

This thesis addresses the statistical integration of metabolomics and epigenetics measurements with gene expression. We approached this data analysis challenge using the Yeast Metabolic Cycle (YMC) as a model system. Gene expression at the YMC can be divided into three, well-defined phases where transcription is coordinated with histone modifications and metabolomics oscillations. First, we analyzed the impact of histone modifications on gene expression, which led to the identification of the histone marks that have a higher impact on gene expression changes. Next, we created a comprehensive, multi-layered, multi-omics dataset for this system by obtaining metabolomics and ATAC-Seq data of the YMC and incorporating an existing nascent transcription (NET-seq) dataset. Moreover, we modeled the impact of chromatin conformation and metabolic changes on gene expression, and created a regulatory model for gene expression, epigenetics and metabolomics by applying PLS Path Modeling, a multivariate strategy suitable for finding relationships across multiple high-dimensional datasets. To our knowledge, this is the first time that PLS-PM is used for the modelling of molecular regulatory layers. We found that gene expression in OX phase was mainly controlled by H3K9ac histone mark and ATP accumulation at this phase, suggesting INO80 ATP-dependent chromatin remodeling activity. We also found an enrichment of H3K18ac during RC phase, together with accumulation of nicotinamide and its derivatives, suggesting that sirtuins may regulate H3K18ac levels at RC to activate fatty acid oxidation response. Aspartate was also associated with RC phase epigenetic regulation, but the mechanisms by which this amino acid may control the epigenome are still unanswered.

Finally, in this work, we have also created Padhoc, a computational pipeline to integrate the existing published knowledge in emerging research fields -such as those studied in this thesis- to propose pathway models that can complement current pathway databases.

Altogether, this thesis involves the generation of a multi-omics dataset that covers metabolic, epigenetic and gene expression information, and their integrative analysis using novel multivariate strategies that model their mechanistic coordination. Moreover, it includes a framework for the reconstruction of biological pathways. All in all, we have presented different strategies by which to study the impact of metabolic changes in chromatin using computational biology approaches.

Resumen

La adaptación celular a ambientes dinámicos constituye un mecanismo esencial para la supervivencia celular. Las células responden a condiciones externas modulando los mecanismos moleculares que regulan expresión génica o la actividad proteica, confiriendo una respuesta rápida a cambios metabólicos externos. Por ello, los mecanismos celulares que captan los cambios metabólicos consistuyen un paso importante en adaptación celular, siendo la epigenética el mecanismo que une el metabolismo con la regulación génica. Las marcas epigenéticas confieren a la célula la capacidad de moldear la conformacion de la cromatina, lo que permite la regulación de la expresión génica. Por tanto, un correcto funcionamiento de la regulación epigenética de la célula, es crucial para la adaptación celular a ambientes con cambios metabólicos.

Los moduladores epigenéticos dependen de la disponibilidad metabólica para poder modificar la epigenética de la célula. Estudios recientes han señalado que la acumulación de ciertos metabolitos es clave para que moduladores epigenéticos actúen sobre las marcas de la cromatina. Un ejemplo claro se ve en los ritmos circadianos, donde los mecanismos epigenéticos median la relación que existe entre las oscilaciones metabólicas y los cambios en expresión génica; la falta de mecanismos epigenéticos desconecta estos relojes moleculares, provocando enfermedades como en el caso del síndrome metabólico. El estudio del control metabólico del epigenoma y el transcriptoma es un área de conocimiento emergente. Muchos estudios han generado información a través de las tecnologías de alto rendimiento, que miden la expresión génica, los metabolitos o las modificaciones de histonas entre otros tipos de moleculas para medir esta conexión, y aunque se ha desarrollado mucha literatura al respecto, los mecanismos que ejercen la regulación de distintos tipos moleculares es todavía desconocida. Una necesidad en el ámbito de la bioinformática es el análisis integrativo de datos moleculares que propongan modelos de regulación detallados para conocer la relación entre metabolismo, cromatina y la transcripción.

En este trabajo se ha aproximado la integración estadística de metabolómica y distintos datos epigenéticos con la expresión génica. Hemos realizado estos análisis integrativos en el sistema modelo del ciclo metabólico de la levadura (YMC), en el cual la expresión génica se coordina con cambios en modificaciones de histonas y oscilaciones metabólicas. Primero analizamos el impacto de las modificaciones de histonas sobre la expresión génica, lo cual nos permitió identificar las marcas de histonas que coordinan los cambios en expresión. Después creamos un conjunto de datos multiómico obteniendo muestras de metabolómica y ATAC-seq en el YMC, e incorporamos un set de datos de NET-seq. Estos datos fueron usados para modelar el impacto de los cambios metabólicos y de la cromatina en la expresión génica y, por primera vez en ritmos biológicos, integramos los tres tipos de datos moleculares en un solo modelo usando PLS-Path Modelling, una estrategia multivariante que permite encontrar relaciones entre muchos conjuntos de datos multi dimensionales. Esta herramienta nos ha permitido conocer que la expresión génica en la fase oxidativa está regulada principalmente por la marca de histona H3K9ac, y la acumulación de ATP en esta parte del ciclo sugiere una regulación de la cromatina activando la enzima dependiente de ATP INO80. El resultado de PLS-PM también nos muestra que los derivados de la nicotinamida podrían afectar los niveles de H3K18ac en a fase RC del ciclo a través de la regulación de las sirtuinas, activando la respuesta de degradación de ácidos grasos. El aspartato también se ha asociado a la regulación epigenética de la fase RC, pero los mecanismos por los que esta asociación novedosa tienen lugar son aún desconocidos.

Finalmente, hemos creado Padhoc, una herramienta computacional

capaz de combinar el conocimiento existente en nuevos ámbitos de investigación -como el de este trabajo- para proponer modelos de redes metabólicas que compleneten el conocimiento de las bases de datos actuales.

Esta tesis recopila la extracción de un conjunto de datos multiómicos que cubre metabolismo, epigenética y expresión génica, así como su análisis integrativo usando estrategias multivariantes novedosas que modelan la coordinación de las distintas moléculas estudiadas. Además, incluimos una herramienta para la reconstrucción de redes biológicas. En conjunto, esta tesis presenta distintas herramientas para estudiar el impacto metabólico en la expresión génica usando la biología computacional.

Resum

L’adaptació cel·lular a canvis dinàmics en l’ambient constitueix un mecanisme essencial per a la supervivència cel·lular. Les cèl·lules responen a les condicions externes modulant els mecanismes moleculars que regulen l’expressió gènica o l’activitat proteica, conferint una resposta ràpida a canvis metabòlics externs. Per això, els mecanismes cel·lulars que detecten els canvis metabòlics constitueixen un pas important en l’adaptació cel·lular. L’epigenètica confereix a la cèl·lula la capacitat de modular la conformació de la cromatina, cosa que permés la regulació de l’expressió gènica. Per tant, un bon funcionament de la regulació epigenètica és essencial per a l’adaptació cel·lular a ambients exposats a canvis metabòlics.

Els moduladors epigenètics depenen de la disponibilitat metabòlica per a poder modificar l’epigenètica de la cèl·lula. Recentment, alguns estudis han apuntat que l’acumulació de metabolits és clau perquè els moduladors epigenètics actuen sobre la cromatina. Un clar exemple són els ritmes circadiaris, on els mecanismes epigenètics intervenen en la relació que existeix entre les oscil·lacions metabòliques i els canvis en expressió gènica; la falta d’aquests mecanismes epigenètics desconnecten aquests rellotges moleculars, provocant malalties com és el cas de la síndrome metabòlica. L’estudi del control metabòlic de l’epigenoma i el transcriptoma és una nova àrea de recerca. Molts estudis han generat informació a través de tecnologies d’alt rendiment, que mesuren l’expressió gènica, els metabolits o les modificacions d’histones entre altres tipus de molècules. Per a estudiar la seua relació és necessària la descripció dels mecanismes moleculars que controlen la regulació metabòlica. Una necessitat en l’àmbit de la bioinformàtica és l’anàlisi integrativa

d'aquestes dades per a proposar els models de regulació que coordinen el metabolisme, la cromatina i la transcripció.

En aquest treball s'ha abordat la integració estadística de metabolòmica i diverses dades epigenètiques amb l'expressió gènica. Hem realitzat aquestes anàlisis integratives en el sistema model del cicle metabòlic del llevat (YMC), al qual l'expressió gènica oscil·la en tres grups de gens, i la seua expressió es coordina als canvis en modificacions d'histones i oscil·lacions metabòliques. Primer vàrem analitzar l'impacte de les modificacions d'histones sobre l'expressió gènica, el qual va permetre la identificació de les marques d'histones que coordinen els canvis d'expressió. Després vam crear un conjunt de dades multiomiques, extraient mostres de metabolòmica i ATAC-seq al YMC, i vam incorporar unes dades de NET-seq. Aquestes dades van ser usades per a modelar l'impacte dels canvis metabòlics i de la cromatina a l'expressió gènica i, per primera vegada als ritmes biològics, vàrem integrar els tres tipus de dades moleculars en un únic model emprant la tècnica del PLS-Path Modelling, una estratègia multivariant troba relacions entre conjunts de dades multidimensionals. Aquest algoritme ens ha permés conéixer que l'expressió gènica en la fase oxidativa està regulada principalment per la marca d'histona H3K9ac, i l'acumulació d'ATP en aquesta part del cicle suggereix una regulació de la cromatina activant l'enzim dependent d'ATP INO80. El resultat de PLS-PM també ens mostra que els derivats de la nicotinamida podrien afectar els nivells de H3K18ac en la fase RC del cicle a través de la regulació de les sirtuïnes, activant la resposta de degradació d'àcids grassos. L'aspartat també s'ha associat a la regulació epigenètica de la fase RC, però els mecanismes pels quals aquesta associació nova tenen lloc són encara desconeguts.

Finalment, hem implementat Padhoc, un algoritme que integra el coneixement existent en àrees de recerca emergents -com les que hem estudiat en aquest treball- per a proposar models de xarxes de regulació que complementen les bases de dades actuals.

Aquest treball recopila l'extracció d'un conjunt de dades multiòmiques que cobreixen el metabolisme, l'epigenètica i l'expressió gènica, així com la seua anàlisi integrativa usant estratègies multivariants que modelen la coordinació de les molècules estudiades. A més, incloem una ferramenta per a la reconstrucció de xarxes biològiques. En conjunt, aquesta tesi presenta distintes estratègies emprades per a estudiar l'impacte del metabolisme en l'expressió gènica emprant la biología computacional.

A Carlos, Pepita,

Voro y Lola.

Gracias por ser los mejores abuelos.

Agradecimientos

Un buen amigo me dijo: ”Una tesis no se termina, se abandona”. Cuando comencé la tesis doctoral tenía la ilusión de cerrar un trabajo de investigación, ahora veo que este trabajo que tanta ilusión tenía por terminar en cuatro años, no ha hecho sino cimentar los inicios de mi carrera investigadora, para la que me han dado herramientas, conocimiento y un sinfín de preguntas. Todo lo que he aprendido se lo debo al trabajo de mucha gente.

Mis ganas de hacer ciencia vinieron del laboratorio de Ciclo Celular de la Universidad de Valencia, donde Inma, junto a Juan Carlos, Mari Carmen, María y Carlos, me enseñaron a ser metódico en mi trabajo, y ante todo tener ilusión. Gracias a mi tía Cristina y el trabajo con Miquel en el laboratorio de David Martínez, decidí tomar el camino de la bioinformática en Wageningen U&R, donde aprendí programación y algoritmos de bioinformática, pero sobretodo aprendí a mirar la biología de otra manera. Esta nueva visión de aproximar los problemas se la debo en gran parte a Dick de Ridder y Marnix Medema, quienes dirigieron mi tesis de master y mis prácticas.

Este trabajo se ha realizado en el marco de una ITN Horizon2020. Pertenecer a una red de doctorado europea me ha permitido conocer a mucha gente en el campo de la metabolómica y epigenética, así como hacer estáncias y visitar distintos laboratorios. En los grupos de Marcus Buschbeck y Andreas Ladurner aprendí el cultivo de líneas celulares gracias al trabajo y apoyo de David y Mehera, mientras que en el laboratorio de Jane Mellor aprendí a manejar el ciclo metabólico de la levadura gracias a Shidong Xi, estancia que ha sido clave para conseguir los resultados que se presentan en este trabajo. Gracias a esta red de doctorado, ChroMe, he vivido experiencias maravillosas en congresos y cursos por toda Europa, aunque lo mejor es la amistad que ha unido a los integrantes de la red de investigación. Hemos viajado juntos y compartido mucha ciencia; Haris, Sana, Iva o Silvia han hecho que este doctorado sea mucho más llevadero. De esta red, me gustaría agradecer al grupo de investigadores que han compartido su conocimiento durante estos 4 años: Oscar, Axel, Catherine o Carles, entre otros, han dedicado muchas horas a nuestra supervisión y formación. Todo el tiempo que han dedicado en nosotros ha tenido un impacto en nuestro trabajo.

Cuando empecé en el grupo de investigación de Ana, sólo estaban Lorena, Cris, Eugenia y Sonia. He tenido la suerte de pasar mucho tiempo con Lorena, que me ha enseñado mucho desde su experiencia y ganas de trabajar, es increíble cómo alguien puede golpear un teclado tan rápido. Pronto llegó más gente al grupo: Fran, Carlos, Ángeles, Manu, Teresa, Pedro... Hemos disfrutado viajes juntos por USA, congresos y, lo que más hecho de menos, las comidas en el CIPF. Muchas personas del ChroMe como Joan, Haris, Laura, Paula o Yang han pasado por el grupo durante estos años y han disfrutado la solidaridad y predisposición de todos mis compañeros/amigos, esta tesis no hubiese sido posible sin vosotros (bueno igual si, pero menudo royo de tesis). Me gustaría mencionar también a la gente del laboratorio de Florida, Raymond, Leandro, Tatyana, Manu, Cecile y Hector. Que hicieron que me lo pasase en grande y aprendiese mucho durante mi estancia allí. También a Victor, Sergio, Anastasiya y Marina, que, ademas de ser unos compañeros geniales, me han ayudado en partes de este trabajo.

Gracias a Biobam por iniciarme al CIPF y por darme cobijo siempre que lo he necesitado. Alex, Marianna, David, Robert, Carlos y Seanna, me habeis ayudado cuando lo he necesitado, me habeis aguantado (solo por eso ya os mereceis estar aquí), además de enseñarme todo lo que pudisteis de la API del B2G, claro. A Stefan, gracias no solo por acogerme en Biobam, sobretodo por ayudarme con todo lo que he necesitado, buscar que estuviese bien y tratarme como uno más durante todo este tiempo.

Agradecer a mis padres todo lo que han hecho por mí, mi educación, compromiso y trabajo viene de su ejemplo y esfuerzo. También a mi hermano y María su apoyo durante todo este tiempo, sin ellos esto no hubiese sido posible. Han sabido manejarme y conseguir despejarme cuando lo necesitaba, además de aguantarme cuando más agobiado he estado. Me gustaría acordarme de alguien especial, mi tía Pepa, que aunque no ha podido estar aquí ha conseguido llegar a mí a través de la gente que la quiere, no puedo evitar pensar que mi camino en ciencia ha empezado gracias a ella.

Me gustaría hacer un reconocimiento especial a Sonia, que me ha enseñado mucho y ha tenido la paciencia de regalarme todos los conocimientos estadísticos que he necesitado en este trabajo; para mí la mejor Co-IP que podría tener, la voy a echar mucho de menos. Y por último y más importante, me gustaría agradecer a Ana todo lo que ha hecho por mí, su apoyo, paciencia, trabajo y ganas de enseñarme, siento que le debo mucho trabajo para hacer honor a todo lo que me ha enseñado. Ana me ha dejado explorar y equivocarme, y me ha transmitido cómo hacer ciencia de forma crítica y constructiva. Muchas gracias.

This work was funded by the following grants:

From 2016 to 2017.

DEANN: Developing a European American NGS Network.

EU Marie Curie IRSES, GA-612583.

From 2016 to 2019.

European Union's Horizon 2020 Research and Innovation Programme.

Marie Sklodowska-Curie Grant Agreement No 675610

National Institute of Health's Common Fund metabolomics program: U24DK097209.

In 2019.

Nuevos métodos para los retos emergentes en el análisis de datos de secuenciación masiva.

MINECO, BIO2015-71658-R

Contents

Glossary

ACL	ATP Citrate Lyase
ANOVA	Analysis of Variance
ASCA	ANOVA–simultaneous component analysis
ATAC	Assay for Transposase-Accessible Chromatin
ChIP	Chromatin Immunoprecipitation
CNV	Copy Number Variation
CQN	Conditional Quantile Normalization
DE	Differential Expression
DEG	Differentially Expressed Genes
DISCO	DIStinct and COmmon simultaneous component analysis
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
FC	Fold Change
FDR	False Discovery Rate
FE	Functional Enrichment
FET	Fisher Exact Test
GEO	Gene Expression Omnibus
GLM	Generalized Linear Model
GO	Gene Ontology
GSE	Gene Set Enrichment
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HMT	Histone Methyltransferase
HOC	High Oxygen Consumption
IS	Internal Standards
JIVE	Joint and Individual Variation Explained
KEGG	Kyoto Encyclopeadia of Genes and Genomes
LISREL	LInear Structural RELationship

LOC	Low Oxygen Consumption
MCIA	Multiple co-inertia analysis
ML	Machine Learning
MORE	Multi-omics REgulation
mRNA	messenger RNA
MS	Mass Spectrometry
NCBI	National Center of Biotechnology
NET-seq	Native Elongating Transcript Sequencing
NGS	Next-Generation Sequencing
NPLS	Multi-Way PLS
ns	not significant
nt	nucleotides
OX	Oxidative
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCR	polymerase chain reaction
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
PLS-PM	Partial Least Squares Path Modeling
PTM	post-translational modifications
RB	Reductive Building
RC	Reductive Charging
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Reads per Kilobase per Million
RT	reverse transcription
SAH	S-adenosylhomocysteine
SAM	S-adenosylmethionine
SE	Sensitivity
SEM	Structural Equation Model
SP	Specificity
SVD	Single Value Decomposition
TF	Transcription Factors
TFBS	Transcription factor Binding Site
TMM	trimmed mean of M values
TPM	transcripts per million
TSS	transcription start site
TTS	transcription termination sites
UniProtKb	Uniprot knowledgebase
YMC	Yeast Metabolic Cycle

Chapter 1

Introduction

1.1 Transcriptional Regulation

Gene Transcription is among the most studied mechanisms in molecular biology. Transcriptional regulation refers to the mechanisms that a cell uses to modulate gene expression to synthesize RNA. RNA is the cellular mediator required for the synthesis of proteins, which are the building blocks that allow a cell to respond to a large variety of intra- and extra-cellular signals and to control important processes such as homeostatic regulation or cell division. The precise control of transcription is critical for the correct functioning of the cell, and hence, gene expression is regulated by a large set of mechanisms that cooperate to ensure cellular viability.

The study of gene expression regulation has been approached from different perspectives in the past 50 years. Positive and negative transcriptional regulation, driven by Transcription Factors (TFs) and post-transcriptional regulation via miRNAs (Figure 1.1A), are among the most well-studied mechanisms, which cooperate to establish the steady-state-levels of transcripts in any given cell type and biological condition [73, 96].

1.1.1 Epigenetics

Epigenetics is a discipline that studies changes in the genome that do not involve alterations of the DNA sequence. These changes are regularly divided into two main categories: nucleotide modifications (e.g. DNA methylation) and histone modifications. While epigenetic DNA modifications are basically restricted to cytosine methylation, histone tails may undergo a large variety of post-translational modifications that include methylation, acetylation and chrotonylation among others [9, 10].

The mechanisms by which histone modifications effect a regulatory role in gene expression are largely unknown and are still under research [45]. So far, the most accepted hypotheses state that chromatin modifications affect the tertiary structure of chromatin, modulating the level of condensation of DNA (Figure 1.1B). The supercoiled chromatin is less accessible for transcription than uncoiled DNA, and histone and DNA modifications change the condensation levels of chromatin by affecting the bonds that pack chromatin [6].

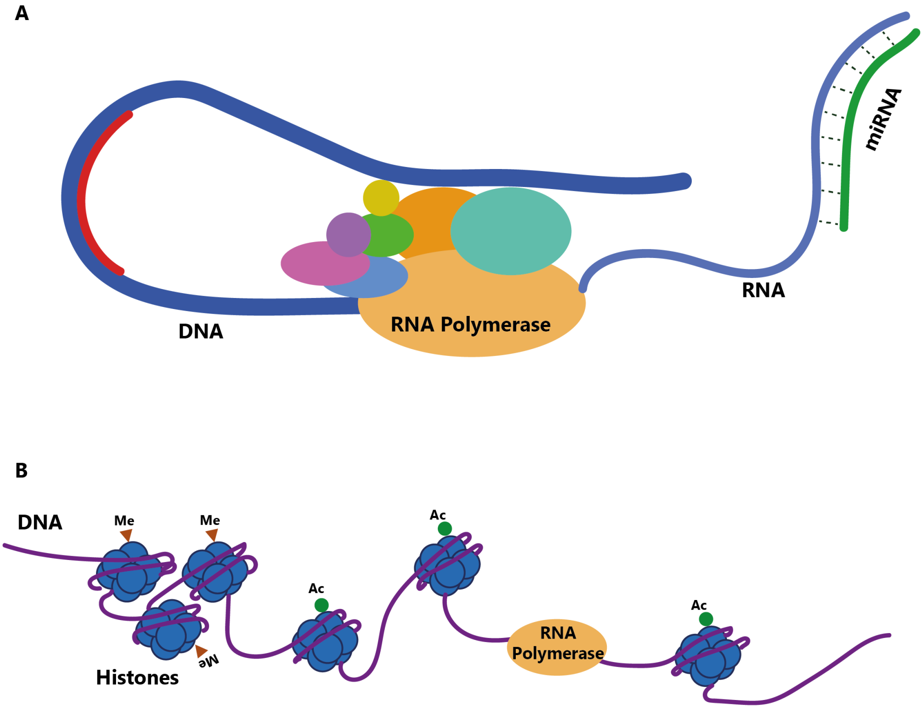


Figure 1.1: (A) Transcriptional and post-transcriptional regulation of gene expression based on TF activity and miRNA. (B) Chromatin conformation affects chromatin accessibility, which offers an extra layer of gene expression regulation.

Methylation, acetylation, chrotonylation and all modifications collectively denoted as *epigenetic marks* involve compounds that derive from cellular metabolism. Enzymatic processes that occur in the cell result in the accumulation of metabolites, such as acetyl-CoA and S-adenosylmethionine (SAM), and different signaling processes mediate the incorporation of these metabolites into chromatin. This implies a connection between the cellular metabolic state and gene expression and the existence of mechanisms by which cells have a transcriptional response to metabolic changes [46].

Table 1.1: A subset of histone modification marks, the genetic position where they distribute in the affected genes and their effect in transcription

Histone mark	Genetic distribution	Transcriptional effect
H3K4me3	TSS	Active/bivalent gene
H3K4me2	TSS	Active/bivalent gene
H3K4me1	TSS	Active/bivalent gene
H3K9me1	TSS and gene body	Active gene
H3K9me2	TSS and gene body	Inactive/bivalent gene
H3K9me3	TSS and gene body	Inactive/bivalent gene
H3K9ac	TSS	Active gene
H3K14ac	TSS	Active gene
H3K18ac	TSS	Active gene
H3K27me2	TSS and gene body	Inactive/bivalent gene
H3K56ac	TSS and gene body	Active/bivalent gene
H4K16ac	TSS and gene body	Active gene
H4K12ac	TSS and gene body	Active gene

1.1.2 Effects of histone marks in chromatin

Histone modifications constitute post-translational modifications of the histone octamer. The most studied histone marks are probably methylation and acetylation, although many others exist, such as chrotonylation or malonylation [179]. Methylation and acetylation occur at lysine residues and, since histones contain multiple lysines in their amino-acid chain, there are many positions susceptible to modification. Additionally, subunits H3 and H4 are those most frequently subjected to metabolic changes (Table 1.1). Modifications imply the covalent bond of the metabolites to the lysine residues, thereby neutralizing its negative chain and resulting in a conformational change in the chromatin. Histones are modified through a balanced coordination of many different enzymes [46].

The groups of enzymes that mediate the modification of histones are generally known as “writers” (Figure 1.2A). Each type of histone mark is targeted by a different subset of enzymes, i.e., Histone Methyltransferases (HMT) for methylation and Histone Acetyltransferases (HAT) for acetylation. The “writers” also include the “erasers”, another subgroup that constitutes enzymes that remove the histone marks with the same specificity as the “writers” (Figure 1.2B). Another group of enzymes that interact with histone marks are “readers”, which bind to modified histones and act as effector proteins (Figure 1.2C).

Histone acetyltransferases (HATs) are the enzymes that catalyze the transfer of an acetyl group to the histone lysine side chains [163], HATs require acetyl-CoA as substrate, and rely on its nuclear availability to operate [112, 198]. The activity of HATs is antagonized by Histone Deacetylases (HDACs), which act as "erasers" of the acetylated histones [46]. These two groups of enzymes (HATs and HDACs) contribute with complementary activities required to establish an equilibrium of histone modifications that control cellular processes in a wide variety of conditions. Chromatin can accumulate a wide range of acetylation marks in different genetic regions, which attract specific readers. This constitutes a histone code that, depending on the combination of modified histones in a specific genomic location, sets the specificity for reader recruitment [45].

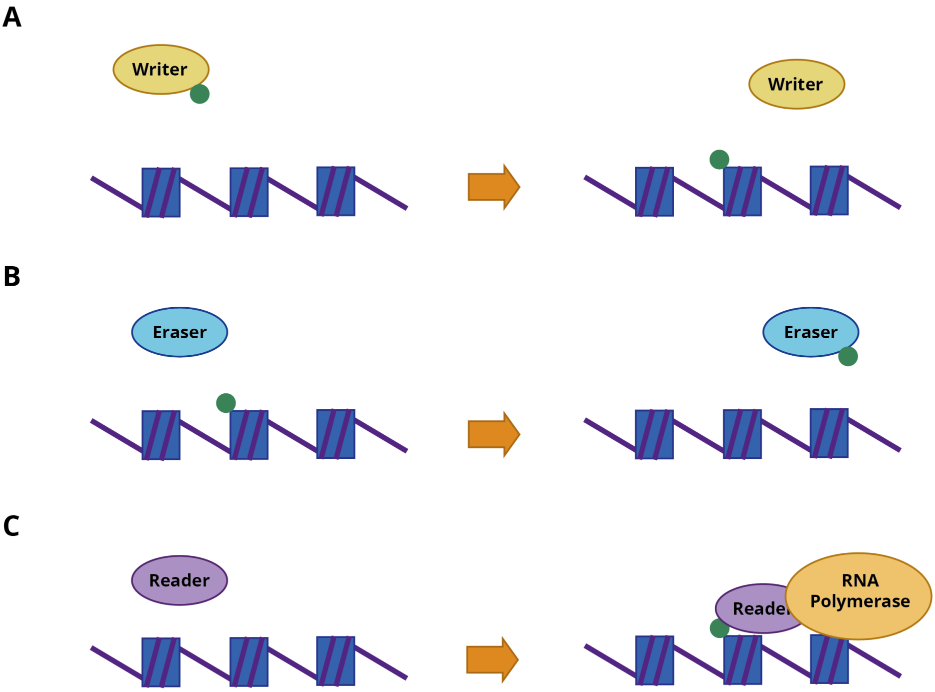


Figure 1.2: (A) Chromatin writers modify histones by adding histone marks to chromatin structure. (B) Erasers constitute special type of writers that remove histone marks from chromatin. (C) Readers bind to modified histones to effect an action on chromatin, for example recruitment of RNA Pol.

1.1.3 Chromatin remodeling

Remodelers are versatile proteins that modify the cell epigenetic landscape by translocating histone octamers from specific DNA regions. Chromatin remodelers are ATP-dependent, and in yeast there are 4 main groups of chromatin remodelers: SWI/SNF, ISWI, INO80 and CHD, all of which share a conserved ATPase core (Figure 1.3) [107, 190].

The SWI/SNF chromatin remodeling complex has a bromodomain that mediates its recruitment to acetylated histones. SWI/SNF has been linked to the regulation of gene expression, both activation and repression, as well as DNA replication and repair. The ISWI family is highly conserved among eukaryotes and plays an important role in high order chromatin organization. Members contain HAND-SANT-SLIDE domains with DNA binding properties [63]. ISWI remodelers change chromatin architecture by altering nucleotide positions, thereby affecting gene expression. INO80 chromatin remodelers are responsible for inositol-responsive gene expression. They bind to histone variants and have been described to be involved in DNA repair and telomere regulation [205]. The CHD remodeling family is comprised of proteins containing two motifs: a chromodomain for DNA-binding and the signature ATPase domain. CHD remodelers have been associated with transcription activation via binding enhancers.

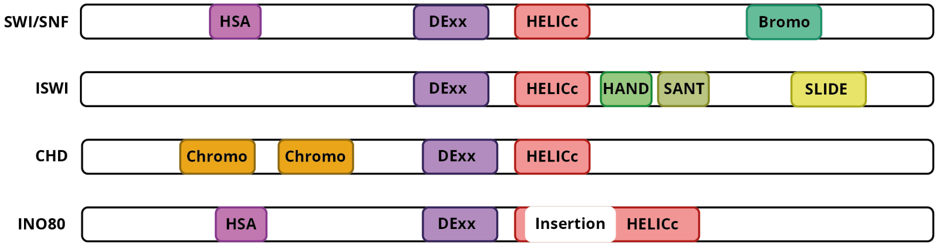


Figure 1.3: Domains in chromatin remodelers. DExx and HELICc domains mediate the ATP hydrolysis; Bromo and Chromo domains bind to acetylated and methylated lysines respectively; HAND, SANT and SLIDE domains recognize nucleosomes and HSA binds to actin-related proteins (Figure adapted from [107]).

Chromatin remodelers are emerging as important enzymes that control the cellular epigenetic landscape. They control chromatin architecture, define open and closed chromatin regions and mediate their accessibility to transcription factors. The coordinated work of chromatin remodelers alongside histone modifiers remains uncertain. Current research in this field indicates that chromatin modifiers operate as the main drivers of epigenetic changes, whereas other research lines highlight the importance of histone marks to facilitate the recruitment of chromatin remodelers [190].

1.2 Metabolic impact on epigenetic marks

Epigenetic marks, especially histone marks, have a dynamic behavior that allows the cell to control specific metabolic and signaling processes. The cell is capable of modulating the marks present in genomic regions where operative genes are transcribed [9, 10], thus regulating their transcription. Their reversible activity allows a flexible response to cellular needs, and the wide range of possible modifications confers a broad combination of cellular signals. The different histone marks require the availability of metabolic byproducts as either cofactors or substrates of the enzymes that catalyze the reaction of histone modification i.e., histone acetyltransferases use acetyl-CoA as substrate to acetylate histones [112, 198].

The cellular mechanisms that connect metabolism, epigenetics and gene expression are relevant to the study of human diseases, such as metabolic diseases and cancer [64]. These complex areas of knowledge cover the contribution of metabolism to the so called “metabolic syndrome” [43]. Understanding the mechanisms of action by which enzymes, chromatin factors and gene regulators coordinate is critical in order to gain new knowledge for the development of therapies for these metabolic diseases. These therapies currently focus on the enzymes that link metabolism and epigenetics [166].

1.2.1 Histone acetylation

Histone acetylation is possibly the best-characterized histone modification. There are at least 13 lysine residues in histones that can be acetylated [111]. Lysine positive charges are neutralized by the acetyl group, which affects the folded chromatin and results in an open chromatin state. Generally, histone acetylation is associated with transcription activation, although the broad range of acetylated lysines contributes to multiple responses, depending on the combinations of histone marks and the external conditions [78].

Histone acetylation is mediated by histone acetyltransferases (HATs). In *S. cerevisiae*, GCN5 and MYST families represent the the main groups of HATs. All HATs require acetyl-CoA as a substrate to acetylate histones, which is an intermediate metabolite in multiple metabolic pathways [104, 111].

Acetyl-CoA is an essential metabolite for anabolism and catabolism. It is a key metabolic intermediate that enters the TCA cycle and is a constituent of structural macromolecules such as lipids. Acetyl-CoA accumulates when a high flux of glucose enters the cell and is processed via glycolysis, or through the degradation of fatty acids via beta-oxidation. Thus, acetyl-CoA can be accumulated both in the presence and absence of glucose [46, 197]. Feeding yeast cells with glucose has been shown to increase ACL-dependent histone acetylation, suggesting that acetyl-CoA availability is a key factor in the activation of this histone mark (Figure 1.4) [197, 198]. Moreover, the accumulation of acetylation marks following glucose uptake occurs at the promoters of growth-associated genes, which links the accumulation of extracellular metabolites with the cellular response to catalyze metabolism of available nutrients.

1.2.2 Histone deacetylation

All histone modifications are dynamic and have conserved mechanisms that mediate their reversibility. Histone deacetylases (HDACs) mediate the removal of acetyl groups from histones, among other proteins. Sirtuins are a special group of well-conserved deacetylases that require NAD+ as a cofactor for their activity [46]. NAD+ is a central metabolic cofactor involved in redox reactions and is required by multiple metabolic pathways from central metabolism; it acts as an electron transfer molecule and alternates between an oxidized state (NAD+) and a reduced state (NADH) [77]. Accumulation of NAD+ is a sign of a low en-

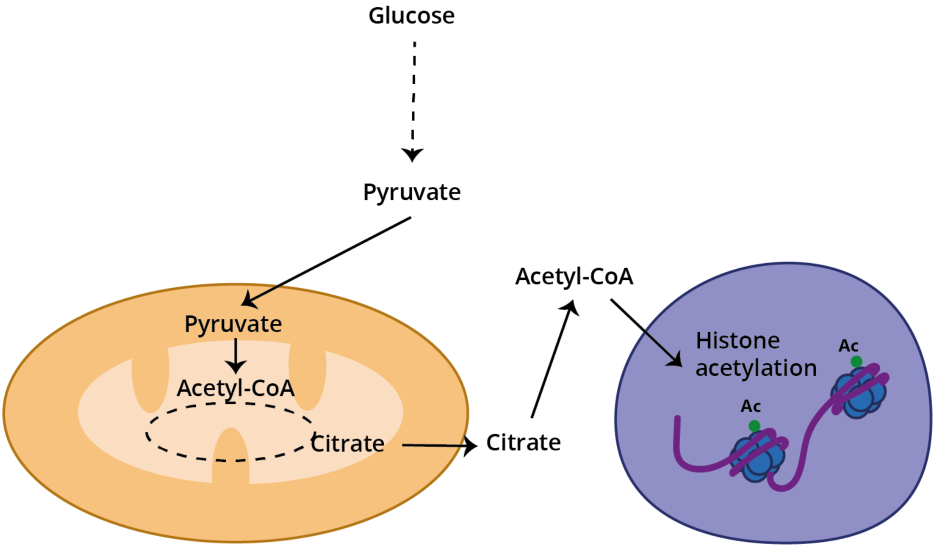


Figure 1.4: Glucose-derived acetyl-CoA can be used by Histone Acetyltransferases to acetylate chromatin, which link metabolic accumulation to epigenetic regulation of gene expression (Adapted from [197]).

ergy state, and sirtuin-dependency of NAD shows that sirtuin activity is nutrient-dependent (Figure 1.5) [53].

Sirtuins have multiple functionalities, including DNA repair, mitochondrial activity or metabolic activity. Their epigenetic and metabolic-sensing capabilities make sirtuins a perfect antagonist to HAT activity [46].

Yeast cells have five types of sirtuins: Sir2 and Hst1-4, while humans have a total of seven sirtuins. Sir2, Hst1, Hst3 and Hst4 are nuclear sirtuins, and each of them have specific acetylated targets. Sir2 has a conserved function in transcriptional silencing and its effects on caloric restriction have been widely studied. Hst1 has been linked to diauxic shift-associated gene repression, while Hst3 and Hst4 play a role in stabilizing the genome during the cell cycle [199].

1.2.3 Histone/DNA methylation

Histone methylation occurs at lysine and arginine residues and has been linked to both activation and repression of transcription, depending on the number of methyl groups attached and the residue modified from the histone tail. Opposite to histone acetylation, multiple methyl groups can be attached to the same

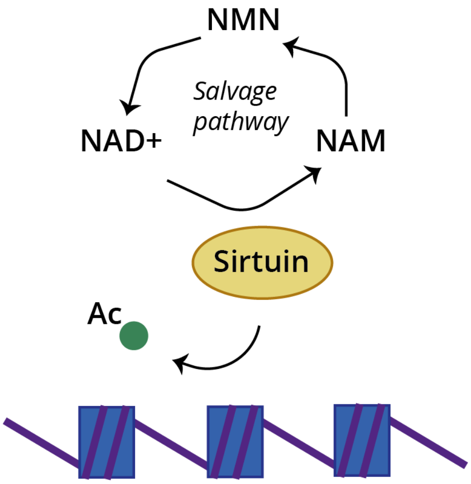


Figure 1.5: Sirtuin activity is NAD-dependent, connecting histone deacetylation with energy metabolism.

histone tail, ranging from a mono-methyl residue to a tri-methylated histone tail. The enzymes that mediate methylation are called histone methyltransferases.

DNA methylation is an epigenetic mark that is generally associated with repression of transcription. DNA methyltransferases methylate cytosines, producing 5-methylcytosine and inhibiting the binding of transcription factors that activate transcription [37, 58].

The intermediary metabolite S-Adenosyl Methionine (SAM) is the methyl donor required by methyltransferases to function [178]. This metabolite is produced from ATP and methionine by methionine adenosyltransferase. SAM levels are maintained in a homeostatic ratio with its byproduct, S-adenosylhomocysteine (SAH), and both are central metabolites in the one-carbon metabolism pathway. SAH acts as an inhibitor of methyltransferases (Figure 1.6) [128].

1.2.4 Histone/DNA demethylation

Methylation is a reversible mark, and demethylases are the enzymes that remove the methyl groups from histone tails. There are two main families of histone demethylases: LSD and JmjC. LSD-driven demethylation requires FAD as a cofactor. FAD is produced from riboflavin and has two balanced forms: FAD+

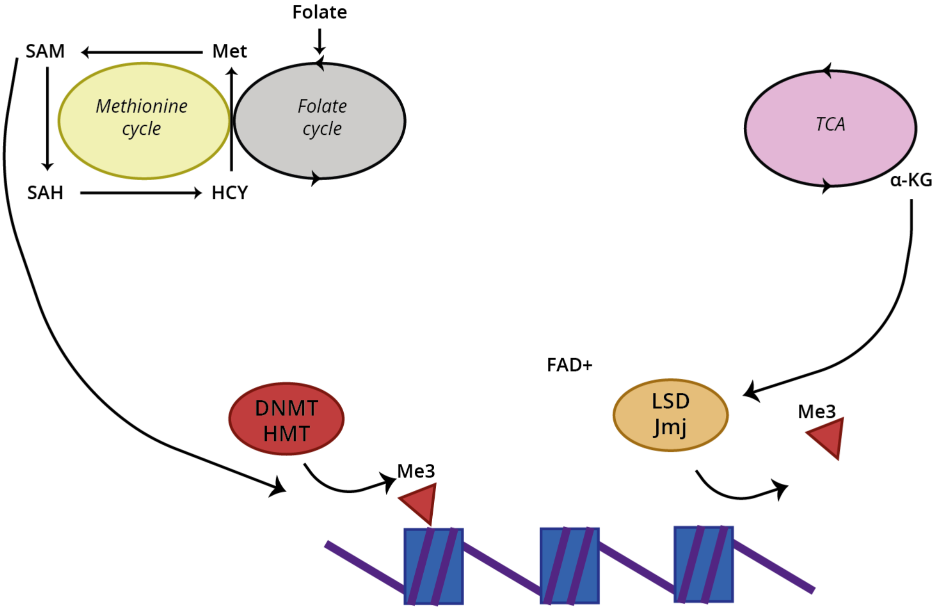


Figure 1.6: Histone methyltransferases are SAM-dependent, and this creates a link methylation with the methionine and folate cycles. Conversely, demethylases are alpha-ketoglutarate dependent (Adapted from [68]).

and FADH. The balance of FAD+/FADH modulates the availability of FAD in the nucleus and its utilization as a cofactor by LSDs [39, 54].

JmjC demethylases have an alpha-ketoglutarate (alpha-KG)-dependent demethylation. This metabolite is produced from isocitrate and acts as an intermediate in the TCA cycle. The role of alpha-KG in modulating the activity of demethylases is still uncertain, but a link between demethylation and the metabolic state of the cell is implicated (Figure 1.6) [39, 80].

1.3 Biological Rhythms

Biological rhythms represent biological processes that have cycling oscillations with variable periods that can last hours or days. The most studied biological rhythms are the circadian rhythms, which constitute daily cycles of gene expression required to maintain cellular homeostasis [60, 152]. The circadian clock allows the regulation of metabolic processes in the day/night cycle, and its malfunction leads to neurodegenerative disorders and metabolic syndrome, among other pathologies [94].

Most studies of circadian rhythms analyze gene expression regulation to identify a master clock regulator that coordinates the oscillation of genes that have a circadian behavior. Chromatin dynamics have also shown circadian oscillations, which suggests coordinated behavior between gene expression and chromatin dynamics [47, 123]. Circadian rhythms can be affected by alterations in the extracellular environment, and this translates into changes in the metabolic state of the cell. Metabolites have been shown to oscillate in a circadian manner. For example, NAD+ has a circadian oscillation that is shown to modulate SIRT1 activity [135]. NAD-dependent SIRT1 promotes histone deacetylation, and several studies have shown that SIRT1-mediated deacetylation controls the transcriptional oscillations of several circadian genes, including Bmal1, Rory, Per2 and Cry1 [135, 149].

The association between metabolism and epigenetic changes raises the question of the mechanisms by which environmental cues can drive gene expression oscillations in circadian rhythms.

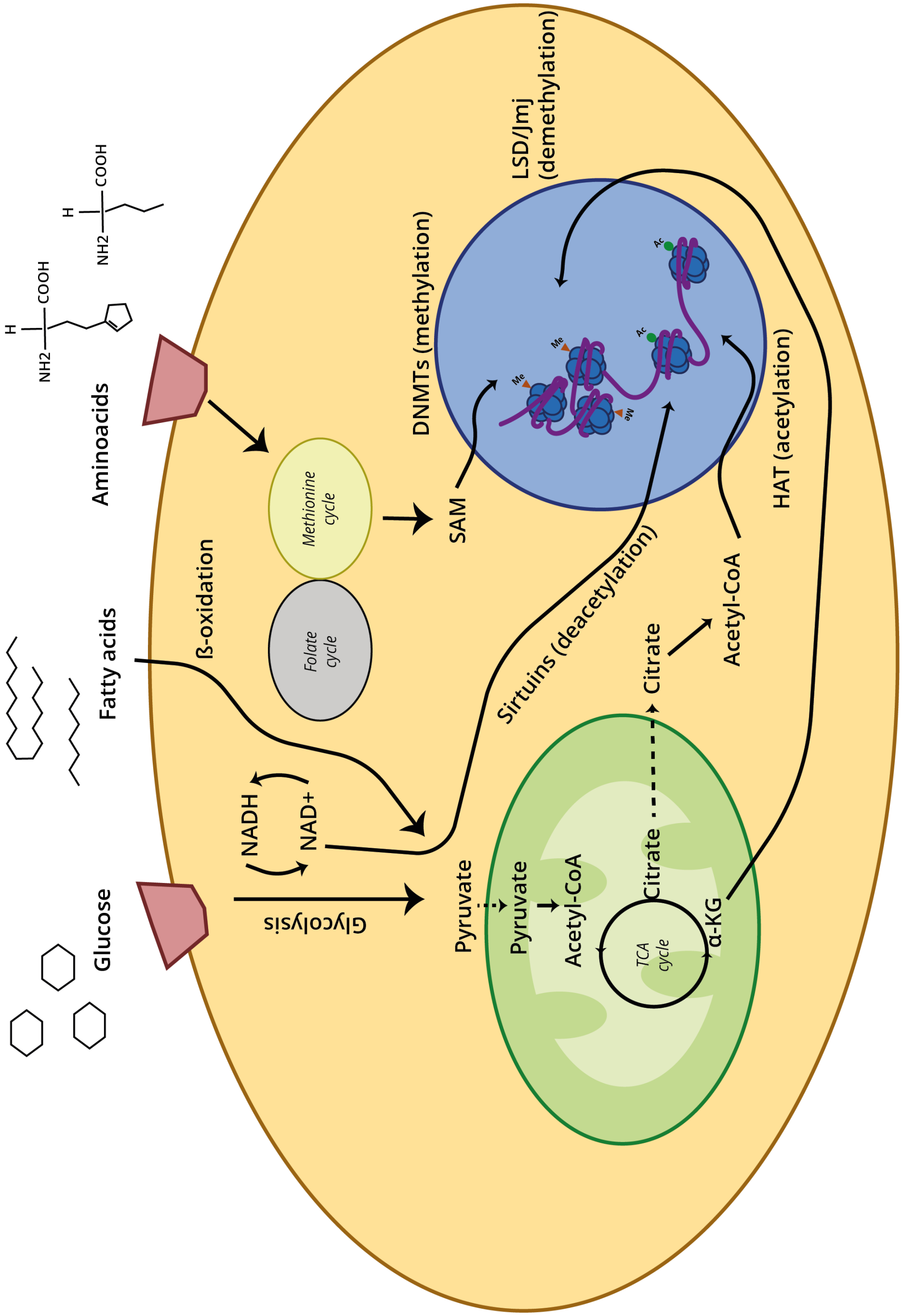


Figure 1.7: Summary of metabolic impact on epigenetic regulation. Combination of histone modifiers with the metabolic reactions that drive the availability of cofactor/substrates to activate the epigenetic mechanisms.

1.4 The Yeast Metabolic Cycle

The Yeast Metabolic Cycle (YMC) is an ultradian rhythm that appears in yeast cells grown under continuous nutrient limiting conditions. Over 60% of yeast genes cycle under these conditions with a robust periodicity. The length of the cycle is determined by the culture flux and is monitored by measuring the level of oxygen dissolved in the media, which reflects the oxygen consumed by the cells [187]. The periodic oscillation in the consumed oxygen suggests that cells shift between two metabolic states, which are usually referred to as high oxygen consumption phase (HOC) and a low oxygen consumption phase (LOC) (Figure 1.8) [126]. The metabolic cycle is synchronized and coupled to the cell cycle, which takes place in a very specific point of the YMC where the proper oxidative state and intermediate metabolites are present [126, 187]. Finally, gene expression in the YMC has been shown to cycle in three clearly differentiated phases where the activity of specific cellular functions is synchronized [102].

1.4.1 The YMC phases

Clustering of the gene expression profiles during the YMC has suggested that transcriptional activity can be divided into three functional phases, which were termed as Oxidative phase (Ox), Reductive Building phase (Rb) and Reductive charging phase (Rc) [187]. Each of the three clusters are formed by different types of genes with specific cellular functions, which confer a very distinctive set of functional properties to each phase (Figure 1.9) [187].

1. **Ox phase** occurs when oxygen consumption starts to increase. The cluster is comprised mainly of genes encoding ribosomal proteins, amino-acids and transcriptional machinery. These functionalities suggest that Ox phase is when most protein synthesis occurs and, since this is a very energetically demanding process, it happens when most ATP is available. Ox phase is the sharpest of the three clusters.

2. **Rb phase** occurs right after Ox phase and ends when the cell lowers its oxygen consumption. Gene profiles from this phase display a lower dy-

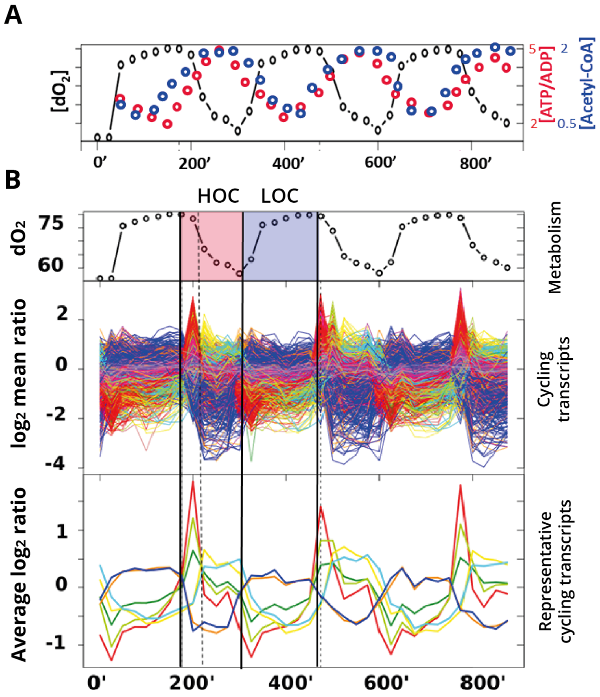


Figure 1.8: The Yeast Metabolic Cycle is characterized by robust metabolic and gene expression oscillations, wherein cells follow synchronous respiration that translates into a repetitive oxygen profile.

namic range. RB phase transcripts are functionally characterized as mitochondrial genes and DNA replication-associated genes.

3. **Rc phase** is the longest of the three phases. Rc phase spans the lowest oxygen consumption interval of the YMC, and functional characterization of genes that integrate this cluster revealed non-respiratory metabolism and protein degradation. This cluster includes genes from the peroxisome, suggesting fatty acid degradation processes occur at this stage of the cycle.

The tight coordination of the three metabolic phases, together with their different functionalities, reveals a well-functioning, sophisticated system with a just-in-time energy supply that traverses multiple metabolic states: the cycle starts by protein synthesis to generate enough cellular resources for metabolic activity and cell division; it continues with a high energy consumption level, via respiratory metabolism, to accomplish cell division; and after this, energy is extracted from lipid reservoirs until there is enough energy to restart the cycle.

The transcriptional regulation of the cycle has been extensively studied since it presumably coordinates the activity of many molecular layers. Transcription factors whose transcripts cycle in any of the three described phases have been analyzed in detail [103]. The study revealed that CRY2 and BMAL1 are possible drivers of the transcriptional changes. However, the study used predicted binding sites from ChIP data; results are limited by this technology, as it cannot provide unbiased chromatin profiling like that of DNAse-seq or ATAC-seq.

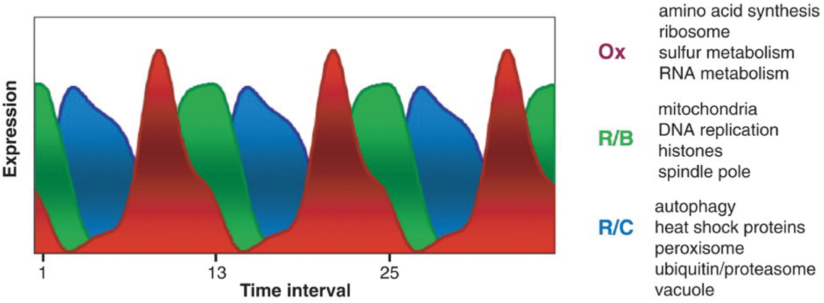


Figure 1.9: Gene expression oscillations have been functionally characterized in three well-defined clusters, that present different functionalities in line with the different cell states (Extracted from [187]).

1.4.2 Epigenetic oscillations in the YMC

Given the nature of the YMC as a metabolically driven rhythm, the changes in gene expression were thought to be regulated by changes in the epigenetic landscape of the cells. After the global acetylation of the yeast cells was confirmed to cycle across the YMC [21], acetylation marks were shown to be essential for YMC progression, since knockout of *gcn5* acetyltransferase prevented cycling. A study from Kuang et al. [102] measured a total of 7 histone marks across the cycle. The sampling time-points were matched to an RNA-seq time-course, and the two datasets were combined to study the mechanisms linking epigenetics with gene expression.

Kuang's study revealed that acetylation signals had more robust oscillatory patterns than methylation signals. The temporal association of the histone marks with gene expression revealed five histone marks that correlate with gene expression changes, including H3K9ac and H3K14ac, the histone marks with an earlier peak than gene expression.

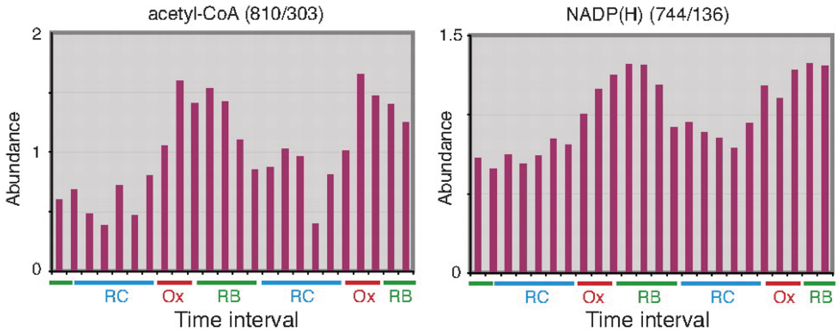


Figure 1.10: Metabolites show an oscillatory behavior across the YMC. Acetyl-CoA (left) and NADP(H) (right) present similar cycling patterns (Extracted from [188])

1.4.3 Metabolic impact in the YMC

Although Kuang's study did not include metabolic information, previous analyses have shown that metabolites have an oscillatory behavior during the YMC. These oscillatory patterns mimic the fluctuations seen in circadian rhythms [186, 188]. Metabolic oscillations result in the accumulation of acetyl-CoA at early Ox phase (Figure 1.10, left), which drives acetylation of growth genes. This directed acetylation can be forced by adding acetate or ethanol to the media, which induces the yeast culture to enter the Ox phase [21].

Metabolomic profiling of the YMC [188] revealed that over 70 measured metabolites cycled in a similar way as that of gene expression, mimicking the transcript clusters. The similarity of cycling profiles between metabolites, histone marks and expressed genes strongly suggests a feedback loop across these three molecular layers and raises questions regarding the detailed mechanisms by which metabolic changes modulate gene expression through epigenetic modifications.

1.5 Omics technologies

High-throughput technologies have been developed to measure the totality of molecules representing one aspect of the molecular biology of the cell. In transcriptomics, gene expression is measured for all genes in the genome; in metbaolomics, levels for all metabolites are quantified; and in epigenomics, reversible modifications of DNA and histones in the chromatin are characterized. Therefore, omic sciences such as transcriptomics, epigenomics or metabolomics are fields of study that benefit from high-throughput techniques to obtain quantitative information from cellular processes. The large amount of data generated by these technologies require powerful computational approaches to analyze and extract relevant information about the cellular processes they measure.

Next Generation Sequencing (NGS) is a high-throughput genomic technology based on the sequencing of DNA and RNA that can be used to measure different aspects of the molecular biology [121]. The most common application of NGS to transcriptome analysis is the measurement of the steady-state levels of messenger RNA, called RNA-seq. However, NGS can also be used to measure other aspects of transcriptional dynamics. For example, techniques such as GRO-seq [33] and Net-seq [24] couple the precipitation of the RNA-polymerase protein to the sequencing of the bound transcripts, thereby providing a measurement of the nascent RNA. Methods such as PAR-CLIP [65] and CLIP-seq [98] evaluate RNA post-transcriptional processes such as transport or splicing, and Ribo-seq measures the amount of RNAs in active translation [44].

Similarly, epigenomic modifications are also measured using NGS technologies. ChIP-seq (Chromatin Immunoprecipitation Sequencing) is a methodology that detects binding sites of DNA-associated proteins [141]. Basically, ChIP-seq starts with the cross-linking DNA to its bound proteins. This is followed by fragmentation and immunoprecipitation of the DNA-binding proteins using specific antibodies. Then, DNA fragments are released and sequenced by NGS. Sequenced reads, when mapped back to the genome, indicate the sites where the protein bound to DNA. ChIP-seq has been widely used to detect binding sites for transcription factors and, when antibodies against specific histone marks are used, can be used to identify the chromatin regions where histone modifications take place. Another recent set of techniques (DNAse-seq, MNAseq-seq and ATAC-seq) measure chromatin accessibility by crosslinking DNA-protein interactions, digesting the DNA that is not protected by bound proteins and then sequencing the protected DNA. These techniques provide information regarding chromatin conformation changes as well as open and closed chromatin states [18].

Mass Spectrometry (MS) has also been used as a high-throughput technique for the analysis of proteins and metabolites. MS measures the mass-to-charge ratio of molecules; it has been used for decades but recent advances in the technology have significantly improved the precision of its measurements. Although MS has been used for longer than NGS, the technology still has a high signal-to-noise ratio, which is especially prevalent in metabolomics studies, as metabolites also rely in different extraction protocols given their diverse biological nature. MS protein detection requires a previous digestion of the protein into small peptides.

1.5.1 Quantification

All these high-throughput technologies provide large amounts of data and require a considerable bioinformatics effort to obtain the actual molecular information. After quality control of the sequenced reads, NGS reads typically need to be mapped to the reference genome or transcriptome of the organism whose DNA/RNA was sequenced [29]. In RNA-seq, reads are mapped to genes or transcripts, and the number of sequenced reads detected for a given gene or transcript indicates its relative expression. ChIP-seq and ATAC-seq follow similar quantification procedures, they require an extra processing step that involves the detection of peaks, which are groups of mapped reads that represent the genomic regions covered by the proteins bound to the DNA [209]. Once peaks are identified above the background noise, they can be quantified using the number of reads that fall within their genomic range. Quantification of omics features is not free of potential biases derived from the sequencing technology and experimental approach. For example, longer genes have more accurate measurements as they accumulate more reads. The GC content of the genomic or transcript region may have an effect on the relative number of accumulated reads. The different total number of reads (or sequencing depth) obtained for different samples makes it impossible to compare them directly, and bias removal and normalization procedures are required to combine NGS measurements for different samples into an homogeneous dataset [29].

MS-derived metabolic recognition occurs after separating the molecules based on a mass/charge ratio. Metabolites are then identified based on their mass/charge ratio and quantified from the intensity of their signal, is basically the area under the spectral peak. Although MS protein quantification is generally achieved using this same procedure, it is also possible to count the number of peptides detected for a given protein, which outputs a similar quantification matrix as that seen for NGS-derived technologies. MS-derived measurements also present extraction and instrumental biases, which are categorized as Type A, when they affect the sample uniformly, and Type B, when they affect individual metabolites differently [81]. Type A biases are generally addressed by the usage of internal standards (IS), but type B biases cannot be corrected with IS, and rely on experimental optimization [81, 131].

1.5.2 Normalization

Once the omic features quantification is obtained, between- and within-sample normalization strategies are usually applied in order to eliminate or reduce technical biases and make samples and features comparable [48].

NGS quantification results in count matrices that contain the number of reads for measured genes/regions in each sample (observation). Samples with higher sequencing depth can lead to unrealistically high quantification values, a bias that is corrected by most of the normalization methods. Other specific biases, like the effect of gene length on expression or the GC content, are mitigated when applying methods such as Reads Per Kilobase Million (RPKM) [133] or Conditional Quantile Normalization (CQN) [67]. Another popular method for

NGS data normalization is the Trimmed Mean of M values (TMM) [153], which accounts for having different distributions of counts across samples.

On the contrary, MS-derived measurements are generally normalized by protein levels in the sample to correct for biases related to the amount of starting material. Normalization methods for metabolic quantification matrices are specific for the distribution of the metabolic signal and include quantile or median normalization, among others [89, 203].

1.5.3 Batch correction

Omic data production is time-consuming and, for large experiments, it is common that samples are not generated simultaneously, but rather in different batches. The economic cost of these technologies also motivates the combination of data from different studies or laboratories, which again translates into different batches. These batches are often an important source of noise in the data that interferes with the biological signal and needs to be removed prior to statistical analysis.

Batch effect correction, which ensures that batch effects are not confounded with a biological effect of interest, is possible as long as the experiment or study has been conveniently designed. Some examples of batch effect correction algorithms are ComBat [113], which uses regression models based on empirical Bayes frameworks to remove noise corresponding to experimental factors, and ARSyN [137], which combines a matrix decomposition based on analysis of variance (ANOVA) with Principal Component Analysis (ASCA model [169]) to extract the signal associated with the experimental factors or with the batch.

1.5.4 Differential analysis

One of the most relevant questions in omics data analysis is identifying the omic features that present changes between experimental groups, across time, etc., known as differentially expressed, or differentially abundant, features. There are many statistical approaches to tackle the differential analysis problem: univariate versus multivariate strategies, parametric versus non-parametric methods, etc.

All of them must deal with the characteristic issues of the omics data, including a low signal-to-noise ratio, reduced sample sizes and a huge number of variables.

The choice of a parametric method depends on the probability distribution of the omics measurements. Traditionally, NGS count data are assumed to follow a negative binomial distribution, while metabolomics data are transformed to meet normality assumptions that can also hold true for NGS data after proper transformations [108]. R packages, such as Limma, implement regression models for normally distributed data, while edgeR [154] or DESeq2 [119] incorporate Generalized Linear Models (GLM) based on the negative binomial distribution. These packages also apply Bayesian approaches to better estimate feature dispersion, given the usually low number of replicates in these datasets and return p-values adjusted for multiple testing since as many tests as the number of omic variables are performed. The maSigPro R package is specially designed to apply GLMs on time-dependent measurements [138]. Among non-parametric methods, the NOISeq R package computes the probability of differential expression by resampling strategies [180].

1.6 Multi-omic studies

The increased availability of omic technologies has created possibilities for simultaneous profiling of different molecular elements in the same biological system and has resulted in multi-omics studies where more than one omic data type is analyzed. While multi-omics experiments offer unprecedented opportunities to model the regulatory mechanisms of the system under study, they also pose new challenges from the statistical point of view. Combining various omic modalities into the same analysis is hampered by differences in dimensionality (e.g., tens of metabolites versus thousands of genes) and the dynamic range or noise level of each one of them [182]. More sophisticated statistical methodologies are needed to address multi-omic analyses, and the biological interpretation of the results is more complicated.

The type of statistical strategy to apply to multi-omics data depends not only on the biological question to answer, and on the nature of the data to be combined, but also on the experimental design, which is an important but often neglected aspect of multi-omics integrative studies. Many integrative analysis methods, especially those based on covariance analysis, require that all the omic modalities are measured on the same subjects (individuals, samples, etc.) [22, 181]. To make this possible, the same biological sample should be able to provide enough material for all omics measurements. Alternatively, experiments should be highly reproducible and allow for confident matching of samples sharing the same experimental conditions. When considering multi-omics studies, these experimental design issues become critical and may determine the type of analysis approach that is applied for data integration.

When the experimental design consists of same or matched samples measured across omics technologies, the integrative analysis of multi-omics data can reveal patterns of common variation across molecular layers that may represent functional interconnections [11, 75] and that can be used to decipher the complex mechanisms that contribute to regulation of expression [71, 143]. There is a variety of statistical models to integrate several omic data types. Here we briefly comment methods based on pair-wise/multiple correlations, strategies based on the analysis of the latent space, and pathway-based approaches.

1.6.1 Correlation-based methodologies

The most straightforward analysis when comparing two omic datasets is to measure the correlation of their features. Correlation-based statistical integration has evolved from simple correlation studies, were two omics are compared and the features that correlate are functionally analyzed, to regression models such as Multiple Linear Regression Models (LRMs) o Generalized Linear Models (GLMs), where a response variable is modeled as a function of a linear combination of predictor variables [136]. While LRMs assume that the response variable follows a Gaussian distribution, Generalized Linear Models (GLMs) offer an extended regression framework that accepts other probability distributions for the response variable, such as Binomial or Poisson distributions. These regression strategies can be used to model the coordination of different omics to regulate the response of another molecule [181]. For example, in a histone modification ChIP-Seq experiment, GLMs can be used to model gene expression changes as a function of the histone changes. If ChIP-Seq measurements of multiple histone modifications are available, the GLM can be constructed as a linear combination of the measured histone marks, which could reveal hidden aspects of the way that histone marks coordinate to control gene expression.

This type of parametric inferential regression model presents some advantages. In general, such models are easily interpretable for non-statistician researchers who are usually familiar with classical regression approaches and interpretation of statistical significance. It is also straightforward to obtain candidates for regulatory elements for specific genes (or other omic features in the response variable). However, when considering several potential regulators as predictors in the model, the multicollinearity problem arises. This problem, together with the usually low sample size in these studies, makes it enormously difficult to have an efficient and robust variable selection. New variable selection strategies such as ElasticNet [210] or Group Lasso [125] regularization methods have been proposed to better deal with these issues, but this is still an open question in the field.

Considering a higher level of complexity, Structural Equation Models (SEMs) are a series of chained regression models in which some variables may act as responses or predictors in different models [191]. SEMs can include both observed and latent (unobserved) variables and need to be fed with a path diagram (a network) connecting the latent variables through potential causal relationships. Inferential SEMs are also known as LISREL (LInear Structural RELationship), one of the most popular software programs to solve them [38]. Despite the potential of SEMs to model causal relations among biological entities, given that a theoretical biological path model can be established beforehand, it is tremendously complicated to apply them to multi-omics datasets for several reasons. Firstly, these inferential models work with strong assumptions about residuals and covariance matrices that are sometimes difficult to meet; secondly, they need a considerable sample size to reliably estimate model parameters.

1.6.1.1 Multivariate strategies

An alternative to the classical correlation-based approaches described above is to use methods based on the analysis of the latent space or multivariate dimension reduction strategies (MVA methods). The most popular methods in this category are probably the Principal Components Analysis (PCA) [16] and the Partial Least Squares Regression (PLS) [59]. While collinearity is an important problem for LRMs or GLMs, dimension reduction strategies take advantage of it. These techniques can also handle the fact that omics data have many more variables than observations. Moreover, MVA methods provide a global picture of all the omic features and samples, efficiently extract signal from noise in these large datasets and require no parametric assumptions on the data. Possible drawbacks of these methodologies might be the difficulty in the generation or interpretation of the models for non-statisticians and the lack of inferential procedures to extract relevant variables. However, they are gaining popularity and there are plenty of tools [155, 184] that provide an easy computation of the models and graphical options to facilitate the interpretation. In addition, some of these tools also include resampling procedures or Lasso variable selection, for instance, to aid in the selection of the most relevant features.

Unsupervised methods such as PCA can be applied on multi-omic datasets by concatenating and weighting the different omic modalities in order to understand the common variation patterns across omics and the relationship among features of different nature [16]. However, there are other possibilities specifically designed for multi-omics analysis. Multiple co-inertia analysis (MCIA) is an exploratory data analysis method to analyze multiple numerical matrices by maximizing their covariance. It has been used for the integration of transcriptome and proteome profiles from NCI-60 cancer cell line, showing a robust selection of features that are used for functional interpretation of the common variability between the two datasets [127]. Other methods that analyze common and distinctive variability include DIStinct and COmmon simultaneous component analysis (DISCO), Joint and Individual Variation Explained (JIVE) or O2-PLS, whose performance was reviewed and compared in [193], and applied to model mRNA and miRNA datasets from glioblastoma multiform brain tumors.

The above unsupervised strategies are applied on two-dimensional matrices. Some experimental designs or datasets allow for N-dimensional structures, for example when different gene-based omics are measured on the same subjects [168]. This would produce a 3-dimensional object with genes in one dimension, samples in the second dimension and omic modalities in the third dimension. For such structures, N-way data analysis is emerging as a methodology by which to reduce the complexity of datasets by projecting the measurements in the latent space. PARAFAC or Tucker3 models can be considered as PCA extensions of these N-way structures. For instance, Tucker3 has been applied to a transcriptomics matrix where gene expression, treatment and time were the three matrix modes, and the decomposition of the three-dimensional matrix allowed to study the interaction between timepoints and treatments and to identify genes with a different time-dependent expression across treatments [27].

Supervised strategies are useful for modelling or explaining the behavior of an omic feature, or a group of omic features, as a function of other features from the same or different omic modality. They can also be applied for classification purposes. The PLS method or PLS variants such as PLS Discriminant Analysis, multiblock PLS, sparse PLS, etc. are some examples of supervised methodologies. PLS Discriminant Analysis (PLS-DA) is an extension of PLS used for classification problems where the response variable is categorical [7] and can be used for sample classification or outcome prediction based on omics features. Multiblock PLS is used in studies where there are multiple omic modalities (blocks) used as predictors of the response matrix [55]. For example, Multiblock PLS could be used to combine several epigenetic omics datasets to predict gene expression as in [117], where copy number variations (CNV), DNA methylation and microRNA were used as explanatory variables for gene expression in a sparse multiblock PLS analysis.

The multiway extension of PLS regression, N-PLS, allows the application of PLS to multifactorial experimental designs where, next to observations and variables, there is a third experimental factor such as time, treatment or different types of measurements of the same set of features [15]. [27] examined the relationship of metabolomics or transcriptomics with physiological variables in pairwise N-PLS analyses, where the three matrices were three-dimensional with common treatments and measured timepoints. A joint interpretation of the results identified several genes and metabolites that were indicators of the physiological changes, and it was concluded that metabolomics was the best estimator of the physiological state.

Finally, there is also a dimension reduction alternative to SEMs, which is the PLS Path Modelling (PLS-PM). As in LISREL models, a path model and a measurement model need to be defined prior to the analysis. PLS-PM assumptions are not so restrictive: it is a distribution-free model, and smaller sample sizes can be used [201]. This makes PLS-PM a more suitable approach to address the analysis of multi-omics datasets, although it is still challenging, given the dimensionality of such data and the difficulty of establishing prior biological path models to be tested.

1.7 Visualization and functional interpretation of multiomic experiments

Cellular metabolism is the set of biochemical processes that occur inside a cell. These reactions involve thousands of proteins and metabolites that produce and consume energy and participate in signaling cascades. Biochemical reactions form an intertwined network in which proteins and metabolites are connected directly or indirectly. Biological pathway databases describe these molecular connections and groups them in functional blocks that represent the existing knowledge of cellular processes. These functional blocks are defined by expert curators that gather information from literature and from experimental evidence to propose a set of finite reactions and pathways that are therefore subjected to human interpretation of molecular relationships. The composition and boundaries of pathways, therefore, may vary from one pathway database to another. Important biological pathway databases are KEGG [85], Reactome

[34], or MetaCyc [87], which describe reactions and pathways mechanistically. Other molecular databases, such as String [174] or IntAct, [70] store information from protein-protein interactions, where each interaction is weighted depending on the level of certainty (computationally predicted have lower weights than experimental discoveries).

Other resources such as Omnipath [189] or NetPath [82] represent signaling pathways, describing cellular processes that are regulated via transduction of a metabolic signal. Gene regulatory factors are elements that interact with chromatin to regulate gene expression, and these regulatory networks are stored in databases such as PAZAR [146] or TRANSFAC [200], among others.

Biological pathways are key to the functional interpretation of high-throughput molecular data. Usually, the analysis of these datasets results in the identification of many significant features (genes, metabolites, proteins), and the biological interpretation of these results proceeds by first looking at the functional properties of the selected features and the biological pathways where they participate. Since this can be a tedious analysis when many features are selected, pathway enrichment algorithms [72, 86, 151] have been developed to find biological processes and functions that are particularly abundant among those selected by the statistical analysis. These enriched pathways are used to study the biological significance of the molecular regulation. Pathway enrichment algorithms rely on the availability of the biological pathway representation to perform the functional interpretations, but unfortunately, there are many biological processes that are underrepresented in pathway databases, especially for non-model species.

In the context of multi-omics data, genes, proteins and metabolites have a direct link to biological pathways, since they form the skeleton of any pathway. Transcription factors are indirectly mapped to genes through their regulatory activity, allowing to detect processes that are transcriptionally activated in the conditions studied. The association of epigenetic elements with biological pathways is less clear than the other mentioned molecules as all genes are susceptible to epigenetic regulation; these regulatory elements can also be quantified, and their presence in chromatin causes gene expression regulation just like transcription factors. The association of epigenetic marks to regulated genes is obtained by determining proximity of the epigenetic mark to the gene regulatory regions. Pathway enrichment algorithms use these direct associations with pathways to perform their overrepresentation analyses.

There are tools that perform joint functional analysis of multi-omics datasets such as the MultiOmics Factor Analysis (MOFA), a computational framework for unsupervised multi-omics integration [3]. Other tools use biological pathways to extract the biological processes that are overrepresented in the differentially expressed features. For example, PaintOmics [72] is an integrative pathway visualization tool that identifies the overrepresented pathways in a set of differentially expressed multi-omic features (e.g., transcriptomics, ChIP-Seq or metabolomics, among others).

In this Introduction we have reviewed the latest discoveries in the metabolic control of gene expression and the different omics and statistical technologies that can be used to study their interaction. Currently, most epigenetic mechanisms that link metabolites with gene expression changes are unknown, and computational approaches that model these complex molecular signals are needed. The development and application of such computational methods are the basis of the objectives for this thesis.

Chapter 2

Motivation, Aims, and Contributions

2.1 Motivation

The cellular mechanisms that perceive metabolic changes to regulate cellular processes are critical for correct environmental response. The study of epigenetic regulation of gene expression has revealed many signaling reactions that connect metabolic changes with gene expression regulatory mechanisms [46, 198]. The enzymes that effect these signaling reactions have been the target of drugs that modulate the impact that nutrients have in the body. Many research groups worldwide study the signaling routes that shape cellular adaptation, searching for novel mechanisms that explain the flexibility that cells display in changing environments.

Although we know that chromatin and metabolism are connected, the molecular mechanisms by which these molecular layers coordinate their activity are still largely unknown. There are many aspects of the metabolic regulation of chromatin that remain unanswered, such as which metabolites impact the signaling enzymes, which epigenetic marks have a stronger effect in metabolic syndrome and how does accumulation of metabolites lead to epigenetic regulation in specific chromatin positions. ChroMe (Chromatin and Metabolism) ETN network was founded to answer these questions. ChroMe consists in the collaborative work of 12 European organizations that host 15 ESRs in a multidisciplinary research project and was created to unravel the mechanisms by which metabolism impacts chromatin. The different ChroMe ESRs tackle the link between chromatin and metabolism from different perspectives, such as studying the glucose-responsive transcription factor ChREBP, the impact of nutrition-derived chromatin changes in physiology or the impact of drugs and lifestyle in disease, among others. Our role as ChroMe partners was to provide bioinformatics software for the integrative analysis of chromatin and metabolism.

Although most research projects that study the nexus between chromatin and metabolism make use of high-throughput technologies to obtain -omic datasets, there is a lack of bioinformatics tools that gather integrative conclusions from multi-omics datasets. Many multi-omics analyses apply omic integration strategies to study how a combination of -omic datasets may explain the variability in another dataset; in other words, the study how two different molecular features contribute to the regulation of a third type of molecular feature [181]. The study of how metabolism affects chromatin and how chromatin, in turn, regulates gene expression implies two consecutive regulatory mechanisms connecting three different molecular layers. This suggests that an integrative multi-omics strategy is needed to capture these two steps, thereby allowing to identify the molecular information that connects the metabolism-epigenome-transcriptome nexus.

Biological pathways are key elements in the analysis and interpretation of -omics datasets; they are especially useful when studying metabolism, as a wealth of information is already available. Biological pathways are stored in pathway databases and represent a curated resource of biological knowledge that contributes to the functional interpretation of the omics data. However, although curated pathways represent reliable descriptions of cellular processes, the pathways' boundaries and elements are often database-specific and do not have flexibility to incorporate new knowledge. This creates limitations for incorporating the emerging knowledge derived from the newest genome research technologies, such as the metabolic control of epigenetic changes, into the pathway data. Therefore, new bioinformatics solutions are needed to quickly update or represent pathway models that keep track of the newest discoveries.

In this thesis, we develop bioinformatics tools for the integrative analysis of metabolism with chromatin and gene expression data. To this end we use existing gene expression and histone modification datasets in the Yeast Metabolic Cycle (YMC), a model system in which metabolites, chromatin and gene expression coordinate oscillatory behaviors in a metabolic-driven system (Chapter 3). Additionally, we generated new metabolomic and ATAC-seq datasets from the YMC to be able to build a comprehensive multi-omics dataset (Chapter 4) that provides enough information of three fundamental molecular layers to study the metabolic-epigenome-transcriptome axis. We model the impact of each -omic dataset on gene expression and develop an integration strategy to model the impact of metabolites on gene expression through changes in chromatin state (Chapter 5). Furthermore, we create a novel software tool to construct *'ad hoc'* biological pathways for biological processes that are poorly represented in the current pathway databases. Our Padhoc tool (Chapter 6) uses text mining to extract the molecular data from the newest literature and combine them with curated, pre-existing information to create biological pathways tailored to the user's needs. With this software, we aim to fill a current gap in pathway analysis, namely, the limitations in applying these methodologies to study the newly discovered biology and pathways from non-model species.

2.2 Aims

The aims of this thesis are:

1) **To understand the impact of chromatin on gene expression across the Yeast Metabolic Cycle.**

   The oscillatory behavior of gene expression in the Yeast Metabolic Cycle has been studied to functionally understand the characteristics of the cycle, revealing a temporal compartmentalization of biological functions. The mechanisms by which this occurs have been associated with histone marks, but the precise transcriptional mechanisms that mediate this control are not yet fully understood. We hypothesize that novel, multi-omics, integrative approaches that combine factor analysis with multivariate regression models can shed light on the interplay between specific transcription factors and chromatin marks to regulate gene expression during the YMC. We do so with the following specific aims:

   - Process the histone marks and gene expression data to obtain matching omic datasets.
   - Weight the impact of each histone mark on gene expression using multivariate analysis tools.
   - Model gene expression changes as a response of histone modifications and evaluate the phase-specific functionalities associated with each histone mark.
   - Infer the Transcription Factors that could coordinate with histone marks to regulate gene expression oscillations.

2) **To extract metabolomics and ATAC-seq data to complete a multi-omics dataset in the Yeast Metabolic Cycle**

   Metabolic impact on chromatin is critical for the regulation of gene expression. Understanding epigenetic oscillations is key to characterizing the Yeast Metabolic Cycle (YMC). However, the current model of metabolic control suggests that oscillatory behavior of metabolism drives epigenetic oscillations. The details of these interactions are largely unknown, and more data on chromatin accessibility and metabolism are needed to answer these questions. Here we aim to obtain “missing” omics datasets of the YMC to be able to create powerful multi-omics models. Our specific aims therefore are:

- Condunct high-resolution, quintuplicate sampling of the YMC to measure metabolic changes.
- Develop a protocol to extract chromatin information from *S. cerevisiae* using ATAC-seq.
- Sample and process YMC cells to apply ATAC-seq to yeast cells across the cycle.
- Process and prepare the in-house (ATAC-seq, metabolomics) and public (RNA-seq, NET-seq and ChIP-seq) -omics datasets by matching the sampling timepoints to create an integrated dataset ready for integrative analysis.

3) **Infer a model of metabolic control of gene expression during the YMC by the statistical integration of the multi-omics dataset.**
The main difficulty in addressing the impact that metabolites have on gene expression lies in the indirect connection between these two molecular layers, which is the dependency on chromatin dynamics to transmit the signal from metabolites to gene expression. This link has been generally studied by integrating metabolomics with gene expression data, but here, we aim to generate a comprehensive model of this process, that includes the dynamic cooperation of three molecular layers to respond to metabolic oscillations. Our specific aims are:

- Use multivariate factor analysis tools (PLS) to model the impact of chromatin and metabolic state in gene expression.
- Use Linear Regression Models to link groups of genes with their potential regulators.

- Use PLS-path modeling to integrate results from previous analyses to propose a global, mechanistic and interpretable model for the metabolic control of gene expression through the impact of metabolites on chromatin state.

4) **To develop a bioinformatics software that creates 'ad hoc' biological pathways.**
Although pathway databases contain a huge number of pathways, recent findings take time to be incorporated in the existing curated biological pathways. In order to provide researchers with a flexible framework to construct biological pathways that include recent molecular discoveries, we propose to develop new software tools that combine curated pathway information with molecular relationships described in the scientific literature. We advance the creation of such software by addressing the following challenges:

- Extract information from pathway databases and combine knowledge within a graph database.
- Obtain scientific articles relevant to the pathway of interest; extract mentions of proteins and metabolites and their embedded relationships as described in the text.
- Create a normalization protocol that merges text-derived information on molecular entities with existing reactions present in pathway databases.
- Extend new pathway construction to non-model organisms by incorporating analysis of orthologues in our tool.
- Create functions to easily navigate new pathway data.

2.3 Main contributions

During this PhD I have contributed to the field of bioinformatics in the form of manuscripts, software, posters and talks. As part of an H2020 ITN I have contributed to several courses, participating as a lecturer to teach -omics data processing, analyses and data integration. I have also actively supervised BSc and MSc theses.

2.3.1 Journal papers

1. Sánchez-Gaya, V.*, Casaní-Galdón, S.*, Ugidos, M., Kuang, Z., Mellor, J., Conesa, A., & Tarazona, S. *Elucidating the role of chromatin state and transcription factors on the regulation of the Yeast Metabolic Cycle: a multi-omic integrative approach.*
   Frontiers in genetics, 9, 578. **2018**.

2. Casaní-Galdón, S.*, Pereira, C.*, Conesa, A. *Padhoc: A computational pipeline for Pathway Reconstruction on the Fly*.
   Bioinformatics, 2020 [in press]

2.3.2 Conferences

- ISMB/ECCB17, 25th Conference on Intelligent Systems for Molecular Biology and the 16th European Conference on Computational Biology. Prague, Czech Republic. July, 2017. ”Automatic reconstruction of metabolic pathways” (Poster).

- ChroMe retreat Conferences. Oxford, UK. July, 2017 ”Developing Bioinformatics methods for the integration of Chromatin and Metabolism data” (Oral Communication).

- Florida Genetics Symposium. Florida, USA. October, 2017. ”Combining databases and text mining for ’on the fly’ pathways reconstruction” (Poster).

- ChroMe mid-term review Conferences. Reus, Spain. April, 2018 “Development of Bioinformatics tools to analyze multi-omics data for the study of metabolic effects on gene regulation” (Oral Communication).

- Bioinformatics@Valencia Meeting. Valencia, Spain. July, 2018. “Combining databases and text-mining for Biological Pathway reconstruction” (Poster).

- Spetses Summer School on Chromatin and Metabolism 2018. Spetses, Greece. August, 2018. “Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status” (Poster).

- Spetses Summer School on Chromatin and Metabolism 2018. Spetses, Greece. August, 2018. “Combining databases and text-mining for biological pathway reconstruction” (Poster).

- JBI2018, XIV Symposium on Bioinformatics. Granada, Spain. November, 2018. “Combining databases and text-mining for biological pathway reconstruction” (Poster).

- ISMB/ECCB19, 27th Conference on Intelligent Systems for Molecular Biology and the 18th European Conference on Computational Biology. Basel, Switzerland. July, 2019. “Combining databases and text-mining for biological pathway reconstruction” (Poster).

- EMBL Symposium: Metabolism meets Epigenetics. Heidelberg, Germany. November, 2019. “Metabolic changes control the epigenetic regulation of the Yeast Metabolic Cycle” (Poster and Flash Talk).

- ISMB20, 28th Conference on Intelligent Systems for Molecular Biology. Montreal, Canada (virtual conference). July, 2020. “A multi-omics approach to characterize the Yeast Metabolic Cycle: using multivariate statistics for -omics integration” (Poster).

- ECCB20, 18th European Conference on Computational Biology. Sitges, Spain (virtual conference). September, 2020. “Padhoc: A computational pipeline for Pathway Reconstruction On the Fly” (Oral communication).

2.3.3 Software

- Casani S, Pereira C and Conesa A.
  Padhoc, platform-independent tool.
  https://github.com/ConesaLab/Padhoc

2.3.4 BSc Thesis Supervisions

- Anastasiya Onofriychuk.
  Role of chromatin modifying enzymes and histone acetylation in the regulation of the yeast metabolic cycle.
  Bachelor's degree in Biotechnology, Polytechnic University of Valencia
  2019

- Marina Villacampa.
  Análisis del estado de la cromatina en el ciclo metabólico de la levadura y su integración estadística con el metabolismo y la expresión génica.
  Bachelor's degree in Biotechnology, Polytechnic University of Valencia
  2019

2.3.5 Master's Thesis Supervisions

- Victor Sanchez Gayà.
  Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status.
  Master's degree in Bioinformatics and Biostatistics, Universitat Oberta de Catalunya
  2018

- Sergio Doria Berenguer
  Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status.
  Master's degree in Bioinformatics and Biostatistics, Universitat Oberta de Catalunya.
  2018

2.3.6 Internship Supervisions

- Sergio Doria Berenguer.

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

  Master's degree in Bioinformatics and Biostatistics, Universitat Oberta de Catalunya

  2018

2.3.7 Teaching

- MIAGE, Multi-omic Integrative Anaylysis of Gene Expression (Centro de Investigación Príncipe Felipe, Valencia). 2017 and 2018 editions, lectures on “Metabolomics”, “Proteomics” and “Hands on Multiomics Integration”.

- Introduction to NGS data analysis: from raw data to intelligible output, COST action. Badalona, Spain. March, 2019.

As part of a European H2020 ITN network, I have participated teaching my fellow ESRs knowledge in bioinformatics and -omics datasets processing and integration.

- ChroMe welcome retreat. IJC, Badalona. December, 2016. “Introduction to Bioinformatics”.

- Statistics for computational biology. ChroMe lectures. Valencia, Spain. April, 2017.

- Specialized seminars on -omics datasets. ChroMe lectures. Stockholm, Sweden. December, 2017

- Multi-omics integration analysis. ChroMe mid-term retreat. Reus, Badalona. April, 2018

2.3.8 Science communication

- Pint of science talk. “Enfermedades metabólicas: la crisis de salud mundial del siglo XXI” May, 2018.
- Act4health – Charity event in Reus coorganized with adc. “Concienciació social sobre la diabetes”. April, 2018

Chapter 3

The role of chromatin in gene regulation in the Yeast Metabolic Cycle

Chapter adapted from Sánchez-Gaya, V.*, Casaní-Galdón S.*, Ugidos M., Kuang Z., Mellor J., Conesa A., Tarazona S. *Joint first authorship.

Elucidating the Role of Chromatin State and Transcription Factors on the Regulation of the Yeast Metabolic Cycle: A Multi-Omic Integrative Approach.

Frontiers in Genetics, 9, 578, **2018**.

3.1 Introduction

The Yeast Metabolic Cycle (YMC) is defined by robust periodic oscillations of gene expression that appear in continuous culture systems under aerobic, glucose-limited conditions. These cycles last about 4–5 hours and exhibit respiratory oscillations alternating between periods of high oxygen consumption (HOC) and low oxygen consumption (LOC). Factors such as glucose concentration and external stimuli can affect period length and amplitude [97, 187]. The nature of the YMC has been extensively studied and is associated with other biological rhythms such as redox cycles and the cell cycle [126]. However, there are still many aspects of the YMC that are unknown or poorly understood due to the complexity of the molecular interactions that coordinate the cellular metabolic state and physiological response during cycling.

Gene expression during the YMC has been characterized using microarrays [97, 167, 187] and RNA-seq [102]. Transcriptional analyses identified three main phases of expression during the YMC: an Oxidative phase (Ox), a Building phase (Rb) and a Charging phase (Rc). The Rb phase was first defined as a reductive phase [187], but recent work [126, 134] has highlighted its oxidative state and proposed it to be part of the HOC phase. Functional profiling of these phases revealed an orchestration of gene expression, which fluctuates in response to environmental conditions, drives the cellular physiological changes and prepares the molecular mechanisms necessary for cycling. Metabolomics studies have shown that metabolite profiles also follow a periodic cycle across the YMC [132, 188], highlighting their importance in enzymatic allosteric regulation and synchronization of yeast ultradian rhythms [126].

Cycling of histone modifications during the YMC confirms that they constitute cellular sensors of the metabolic conditions. For example, cycling levels of acetyl-CoA (cofactor for histone acetylation) reflect alternative high and low energy states of yeast cells [21] and might be key to coordinating gene expression [102]. Kuang and co-authors showed that chromatin changes have a temporal association with transcripts, as both present similar oscillations. They studied the correlation of gene expression clusters with histone modifications to reveal the contribution of each histone modification to the regulation of the different YMC phases; furthermore, they showed a sequential regulation of genes involved in transcription, mitochondrial activity, cell cycle and different metabolic processes along the YMC. However, in their study, no significant relationships were established between histone modifications and the expression of specific genes or transcriptional regulators, leaving unanswered questions regarding the functional orchestration of the system.

Few studies have investigated the potential role of transcription factors (TFs) in the regulation of the YMC. Rao and Pellegrini [150] analyzed the periodic activities of TFs to explain the regulation of the YMC phases, while Kuang [103] inferred the spatio-temporal DNA binding of important TFs across the cycle. While it is reasonable to assume that regulation during the YMC consists of a combination of histone modifications coupled with TF control, these two aspects have never been jointly studied. A combined analysis of these datasets would facilitate understanding of the contribution of each regulatory layer to the transcriptional dynamics observed in the YMC, and to decipher the significance of the interaction between specific histone marks and TFs in controlling gene expression.

In order to shed light into the regulatory mechanisms behind the YMC, we present here a novel strategy for the integrative analysis of the chromatin state and gene expression in this process. We used data from Kuang [102], which contains ChIP-seq experiments for 8 different histone modifications and matching RNA-seq data. In addition, we included a ChIP-seq dataset on an additional histone modification (H3K18ac), which turned out to be a key regulator of YMC. In this chapter, we analyzed the interplay between chromatin status, transcription factor binding and gene expression, and identified a core set of TFs that are relevant to the synchronization between histone marks and transcriptional regulation. Our results indicate that histone modifications contribute differently to the YMC progression, and we identified several TFs that might participate in the molecular regulation of the cycle. Overall, the integrative analysis described in this chapter unravels regulatory mechanisms controlling switches in cellular processes that allow yeast to respond to factors affecting the metabolic cycle.

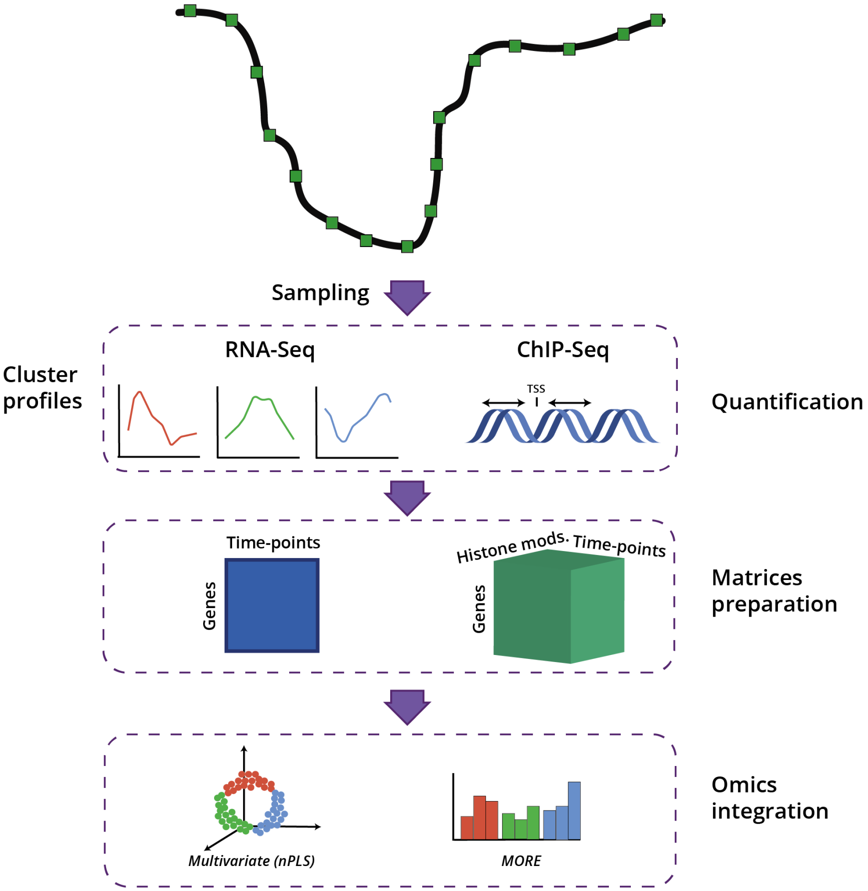


Figure 3.1: Chapter 1 analysis workflow. The Yeast Metabolic Cycle was evenly sampled to obtain matching timepoints for RNA-Seq and histone modifications. In this work we quantified the RNA-Seq and ChIP-Seq datasets and prepared them for an integration analysis in which we modeled the regulatory role of chromatin on gene expression using PLS and GLMs.

3.2 Methods

3.2.1 Omics Data Sets

Experimental design

Gene expression and histone modification data from Kuang et al. [102] were retrieved from the Gene Expression Omnibus (GEO) repository (accession number GSE52339). We also included an additional histone modification ChIP-seq experiment for H3K18ac, provided by Dr. Mellor's laboratory, with GEO accession number GSE118889. Kuang's dataset was complemented with H3K18ac ChIP-seq, given its relevance in transcriptional regulation [35, 196]. All omic measurements were obtained from YMC experiments as described in Tu et al. [187]. As the duration of the different phases of the YMC is not uniform, samples were unevenly taken in each phase to generate an equal number of time points for all phases of the cycle [102]. The number of sampling points was 16 for both RNA-seq and histone modification data.

Gene Expression

Gene expression was measured by conducting RNA-seq on an Illumina HiSeq 2000 platform to generate single-end, 50 bp-long reads. Data were pre-processed as in Kuang et al. [102]. Basically, expression data were normalized by sequencing depth, log transformed and centered per gene.

Histone modifications

Histone modifications were measured with ChIP-seq using antibodies against eight different marks: H4K16ac, H3K36me3, H3K4me3, H4K5ac, H3K9ac, H3-K56ac, H3K14ac, H3K18ac. H3 was used as a control. Ten biological samples per time point were obtained in two different batches, H3K9ac was measured in both. The samples were sequenced using three different technologies (Illumina HiSeq 2000 ChIP, Illumina Genome Analyzer ChIP, and AB SOLiD System), which provided read lengths from 35 to 51 bp. ChIP-seq data were processed as detailed in the next section.

3.2.2 ChIP-Seq Data Processing

3.2.2.1 Quality filtering

The quality of ChIP-seq fastq files was assessed using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were quality-filtered and trimmed to discard low quality reads before mapping to the reference genome. The Trimmomatic software v0.32 [14] was applied with restrictive filters on the highest quality sequencing samples. Minimum quality was set to 30 (restrictive) or 25 (non-restrictive) at the beginning and end of the read, with a sliding window of 5 and a minimum length of 28 bp (restrictive) or 25 bp (non-restrictive). H3K36me3 and H4K16ac histone modifications were discarded due to their low sequencing quality.

3.2.2.2 Read mapping

Sequencing reads for the remaining histone modifications were mapped to the Ensembl database (release 91) *S. cerevisiae* reference genome [206] using Bowtie [106]. Multi-mapped reads were discarded. For AB SOLiD reads, the reference genome was first converted to color space coding using bowtie-build with parameter –c. Duplicated reads were removed with Samtools [115]

3.2.2.3 ChIP-seq Quantification

In order to obtain ChIP-seq quantification values, coverage per nucleotide was calculated for the whole genome with the program genomecov from the bedtools suite [147], specifying the parameter –d. After an exploratory analysis of the coverage across different genomic regions, we defined two genomic regions per histone modification and gene that corresponded to either the 300 bp upstream or downstream regions from the gene transcription start site (TSS). For each gene and region, the average coverage was computed using in-house Python scripts and the yeast genome annotation (gtf file) from Ensembl (release 91) [206]. Consequently, two quantification matrices were generated for each histone modification, one for upstream and one for downstream from TSS (Figure 2).

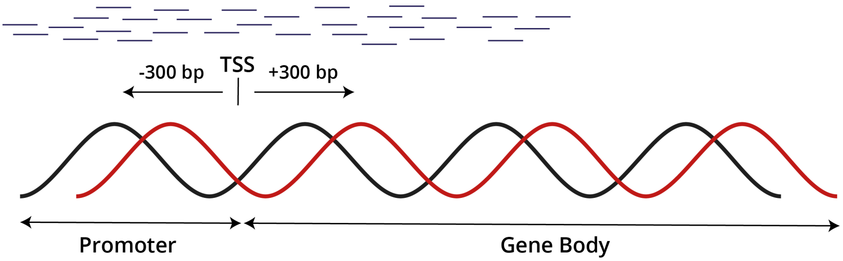


Figure 3.2: Sketch of the position of the promoter and gene body in a gene, and the regions that were quantified in our ChIP protocol, which corresponded to 300 bp upstream and downstream of the TSS.

The H3 ChIP-seq data were used as a control for normalization of quantification values. Each histone modification sample was corrected by its sequencing depth, divided by the H3 sample matrix, log-transformed and centered. Centering the data matrix was enough to correct for batch effects as validated using the ARSyN package. After this correction, the H3K9ac dataset with the highest quality was selected for further analyses.

3.2.3 Gene expression analysis

3.2.3.1 Differential expression

Differential Expression analysis of RNA-seq data was performed with the R maSigPro package [26, 138], which applies a polynomial regression model to analyze time-course gene expression data (equation 3.1). A polynomial degree of 3 was selected, as it provided the model with the highest adjusted R^2. Differentially expressed genes (DEGs) were called by having a significant model (False Discovery Rate (FDR) adjusted p-value $<$ 0.05) and a minimum R^2 value of 0.6.

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon \tag{3.1}$$

Where y are the expression values of a given gene, β are the model coefficients, t is time and ϵ the error term.

3.2.3.2 Clustering and silhouette analysis

DEGs were clustered within maSigPro using the k-means method and a total number of 3 clusters. The silhouette coefficient [156] was used to measure the quality of clusters. Silhouette assesses the quality of the clustering results by comparing the distances between elements in the same cluster to the distances between different clusters. Elements of a cluster with a low Silhouette coefficient are likely to be wrongly classified. We used the silhouette plot to determine if gene expression profiles had been well clustered using our Differential Expression and clustering pipeline.

3.2.4 Time-point alignment

Histone modifications and gene expression data were obtained from the YMC at 16 sampling time-points and two different experimental runs. As samples were unevenly distributed in the two data extractions, we took advantage of the monitoring of the Yeast Metabolic Cycle progression using the dissolved oxygen levels and aligned the samples according to their position in the cycle; timepoints 10 and 11 from the RNA-seq data, and timepoints 13 and 14 from histone modification data were averaged, resulting in two matrices with matching observations. Details of this procedure are shown in Results section 3.3.1.

3.2.5 Multi-way Partial Least Squares Regression (N-PLS)

Partial Least Squares regression (PLS) is a multivariate regression method commonly used to study the relationship between a response matrix and a predictor matrix [59, 159]. These two matrices typically contain observations (e.g., time points) in rows and variables (e.g., genes) in columns. PLS obtains a set of new variables (components) that are a linear combination of the original variables and recapitulates most of the covariance between the response and predictor matrices. PLS is, therefore, a dimension-reduction technique that allows for summarization of the relationship between two high-dimensional data structures.

PLS models the relationship between two, two-dimensional matrices derived from studies with two experimental factors. The generalization of PLS to data structures with additional experimental factors is provided by N-PLS. We used the N-PLS framework described in [15] and adapted in [17] to accommodate the experimental factor represented by the type of histone modification. In N-PLS, the dimensions are called modes. Therefore, in our case, we have two modes for $\mathbf{Y}$ matrix and three modes for $\underline{\mathbf{X}}$ matrix: where the first mode captures the genes, the second mode captures the time-points and the third mode refers to histone modifications (Figure 3.3). We used the N-PLS framework to model Gene Expression as a function of the histone modification signal at two genomic regions: 300 bp prior TSS and 300 bp after TSS.

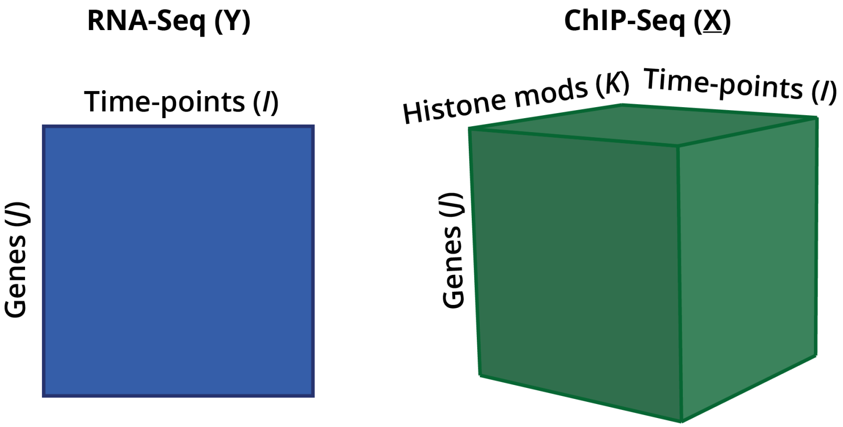


Figure 3.3: N-PLS is an extension of PLS that allows for the usage of three-dimensional matrices. In our study, we used N-PLS to model gene expression based on a three-dimensional histone modification matrix, where we included genes, timepoints and the histone modifications as the three matrix modes or dimensions.

In N-PLS, as in PLS, scores and loadings represent the projection of the observations and variables, respectively, into the latent space. Additionally, in N-PLS, scores and loadings are calculated iteratively and projected in the same latent space. Ultimately, each mode can be separately represented in the lower dimensions as in PLS. However, in N-PLS, an additional element is computed, the core matrix, which indicates how the components of each mode are associated. For example, the first component of mode 1 (in our case genes) may be linked to the second component of mode 2 (in our case time) and the first component of mode 3 (in our case histone type), resulting in a core element (1,2,1). Moreover, each core element has an associated explained variability that can be used to evaluate the quality and interpretability of the model.

Using $\underline{\mathbf{X}}$ $(I{\times}J{\times}K)$, and its unfolded version $\mathbf{X}$ $(I{\times}JK)$; and $\mathbf{Y}$ $(I{\times}J)$; the N-PLS algorithm calculates the latent spaces W^J and W^K that maximize covariance between $\underline{\mathbf{X}}$ and $\mathbf{Y}$. $\mathbf{X}$ (equation 3.2) and $\mathbf{Y}$ (equation 3.3) can be written using the PLS expressions from [17]:

$$\mathbf{X} = \mathbf{T}\underline{\mathbf{G}}(\mathbf{W^K} \otimes \mathbf{W^J})^T + \mathbf{R_X} \tag{3.2}$$

$$\mathbf{Y} = \mathbf{U}(\mathbf{Q^J})^T + \mathbf{R_Y} \tag{3.3}$$

Where $\mathbf{T}$ and $\mathbf{U}$ are the loading matrices from $\underline{\mathbf{X}}$ and $\mathbf{Y}$, respectively, $\mathbf{W}^K$ and $\mathbf{W}^J$ are $\underline{\mathbf{X}}$ weight matrices and $\mathbf{Q}^J$ is $\mathbf{Y}$ weight matrix. $\underline{\mathbf{G}}$ is the core matrix, which explains the association between the components, and $\mathbf{R_X}$ and $\mathbf{R_Y}$ are the residual matrices from $\underline{\mathbf{X}}$ and $\mathbf{Y}$, respectively. An iterative process based on SVD is applied to obtain $\mathbf{T}$ and $\mathbf{U}$, whose covariance is maximized. $\mathbf{T}$ and $\mathbf{U}$ are linearly related as in equation 3.4, where $\mathbf{B}$ is the coefficients matrix and $\mathbf{R}_U$ are the residuals.

$$\mathbf{U} = \mathbf{TB} + \mathbf{R_U} \tag{3.4}$$

3.2.6 MORE Regression Models

Gene-specific regulatory models were built using the MORE (Multi-Omics REgulation) package (https://github.com/ConesaLab/MORE). MORE fits a Generalized Linear Model (GLM) to explain gene expression as a function of the levels of regulatory elements $R_1, ..., R_p$ (Equation 3.5). Here, we used transformed expression data for which we assume normality. Starting from a set of gene-specific regulatory elements containing cis- (i.e., promotor histone methylation) and trans- (i.e., putative promotor-binding TFs) features, MORE first applies variable regularization based on low variability and multicollinearity. Secondly, two different variable selection procedures are applied to obtain a regression model for each gene: Elastic Net penalized regression [211] and stepwise regression [41]. Features with significant model coefficients are considered to be active gene expression regulatory elements of the gene expression.

$$y = \beta_0 + \beta_1 R_1 + \beta_2 R_2 + ... + \beta_p R_p + \epsilon \tag{3.5}$$

Where y is the gene expression, β_i are the model coefficients and R the omic regulators. In this study, we created gene-wise MORE models using either transcription factor expression or histone modification data as regulators ($R_i, i = 1, ...p$), as described next.

3.2.6.1 Histone modification regression models

In the first MORE model, histone modifications were used as explanatory variables. We considered two regions defined for each of the six histone modifications considered, resulting in an initial model that included 12 regulatory variables for each gene. Genes with a significant (p-value $<$ 0.05) coefficient for at least one histone mark were selected as displaying active chromatin regulation.

3.2.6.2 Transcription Factor Regression models

The second MORE model considered differentially expressed TFs as predictors. The potential target genes of TFs were retrieved from the Yeastract database [183]. After model adjustment, significant regulators of genes were selected by having regression coefficient p-values $<$ 0.05. The final result of this analysis is a gene-TF association matrix indicating significant regulation of the gene by the TF.

3.2.7 Functional Enrichment Analysis

Functional Enrichment analysis was applied to the set of genes with a significant coefficient in each regulatory factor (TF or any histone mark) as obtained from MORE models. Functional terms were retrieved from Gene Ontology (GO) [4, 31] and from Kyoto Encyclopedia of Genes and Genomes (KEGG) [83, 84] databases, and the Fisher's Exact test was applied to select the significantly enriched functional categories (p-value < 0.05). The genes regulated by each histone modification were grouped according to the YMC cluster they belonged to, and the functional enrichment analysis was performed on each of the three groups. For TFs, the number of regulated genes was much lower, and the enrichment analysis was done without the separation of the YMC phases.

3.2.8 Selection of the Most Relevant Regulators

An enrichment analysis approach was applied to select the most relevant regulators during the YMC. We considered a TF to be relevant if the proportion of MORE significant models involving this TF was significantly higher than the proportion of the TF regulations in Yeastract database. Additionally, we selected most relevant histone modification-TF pairs as those co-regulating a proportion of genes higher than expected by chance. Both analyses were done using the Fisher's Exact test (FDR adjusted p-value < 0.05 and odds ratio > 1).

3.3 Results

3.3.1 Time point alignment

Kuang et al. [102] measured gene expression and histone modifications at 16 sampling points, as shown in Figure 3.4A. However, the sampling scheme was not fully coincidental. For instance, the RNA-seq RB phase had 7 time-points, while ChIP-seq only included 6 time-points. Since our omics integrative analysis requires matchable observations, a time-point alignment was necessary. We used oxygen consumption levels as a cellular metabolic state indicator to compare the two experiments and match disagreeing time-points. As a result, time-points 10 and 11 from RNA-seq were averaged, as well as time-points 13 and 14 from ChIP-seq, resulting in final data matrices with 15, rather than the original 16 time points (Figure 3.4B).

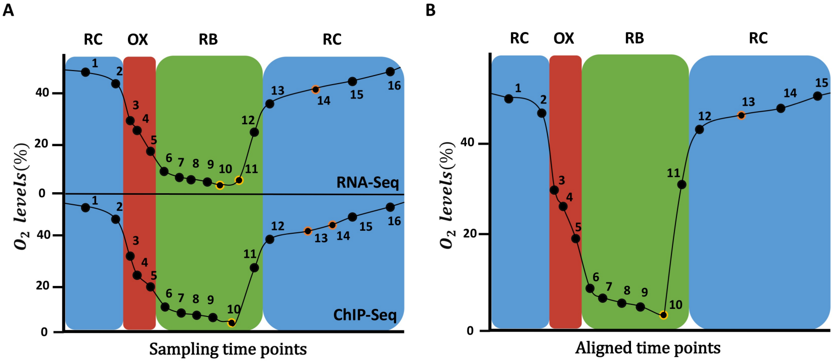


Figure 3.4: Time points sampled during the YMC. (A) The original time points sampled for RNA-Seq data (top) and ChIP-Seq data (bottom) at each YMC phase are shown. (B) The 15 time points after the alignment of the two time series are displayed: time-points 10 and 11 from RNA-seq were averaged, as well as time-points 13 and 14 from ChIP-seq. On the Y axis the percentage of oxygen in the environment is indicated.

3.3.2 Omics capture YMC variability

3.3.2.1 RNA-seq

RNA-seq was subjected to a PCA exploratory analysis, which revealed that gene expression variability followed a circular pattern (Figure 3.5A). Gene expression samples did not only group by their YMC phase, but also captured the temporal differences between samples of the same phase. This indicates that YMC temporal variability is present in the RNA-seq dataset.

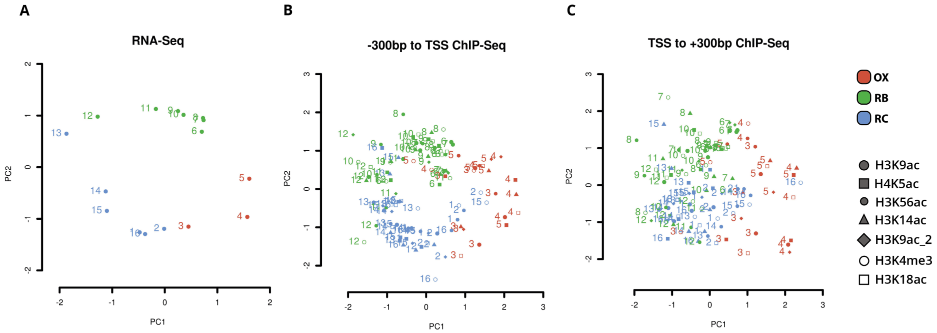


Figure 3.5: Principal Component Analysis (PCA) on gene expression (A) and histone modifications using the 300bp upstream of the TSS quantification (B) and the 300 bp downstream of the TSS quantification (C). Histone modifications PCA includes all histone modification used for this study.

3.3.2.2 Histone Modifications

Similarly, PCA exploration of histone modification levels indicated that samples followed a similar trend as in gene expression for both regions analyzed. The first two PCA components discriminated YMC temporal differences, with histone mark signals at promoter regions resulting into a better separation of the YMC phases in all evaluated histone modifications (Figure 3.5B and C).

Gene Clustering into YMC Phases

maSigPro analysis identified 2,552 genes with significant expression changes across the YMC. PCA of these DEGs showed clear separation of the time points according to the different metabolic phases of the YMC (Figure 3.5A), which corroborated the good quality of the data and of the differential expression results.

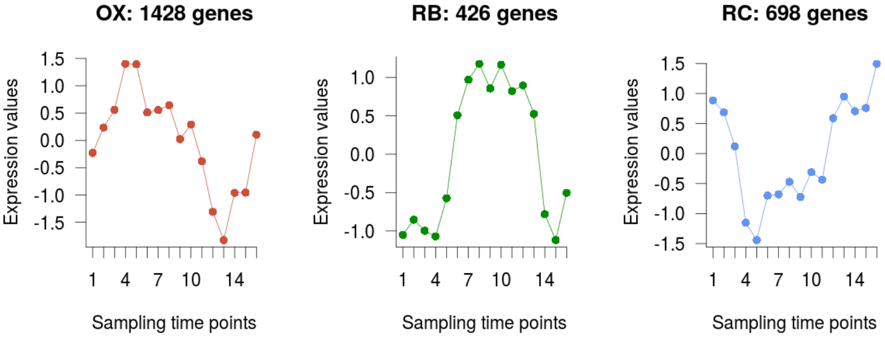


Figure 3.6: Gene clustering. Average profile of the differentially expressed genes in each of the three clusters obtained, which are in turn associated to each YMC phase.

maSigPro also clustered these 2,552 DEGs into three groups that recapitulated the YMC phases [102, 187]. The number of genes assigned to each of the clusters was 1428, 426 and 698 for OX, RB and RC respectively (Figure 3.6). Clusters were assigned to their corresponding phases by localizing the position of the expression peaks from the gene profiles in the time-course. We assessed the quality of our clustering results with the Silhouette metric and compared them to the previously reported clusters in Kuang et al. [102]. While the clusters identified in our study and in Kuang et al. [102] largely agreed (Figure 3.7), a better clustering performance was obtained with our analytical strategy (Figure 3.8).

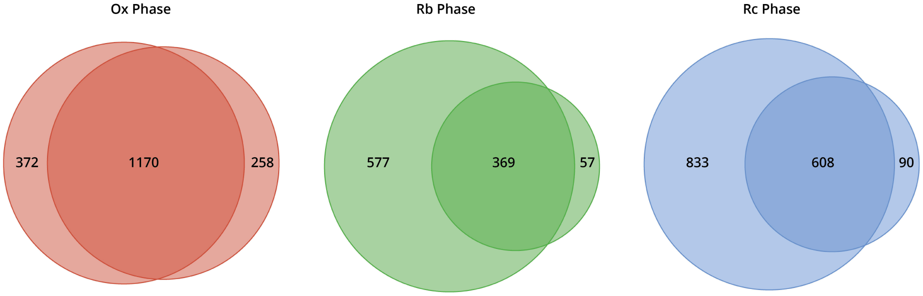


Figure 3.7: Overlaps between the YMC clusters defined in Kuang (2014) and our paper using maSigPro models of the three YMC phases.

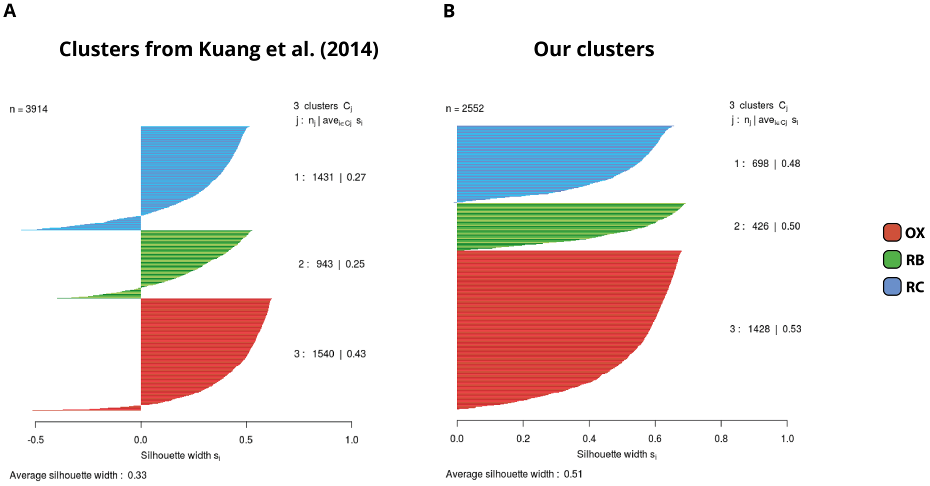


Figure 3.8: Silhouette plot of the gene clusters from this project (A) and from Kuang et al. (2014) (B). Clusters represent RC genes (blue), RB genes (green) and OX genes (red).

Expression changes in the YMC were mostly driven by H3K9ac and H3K18ac

N-PLS was applied to explore general relationships between gene expression and chromatin status of the DE genes. Note that RNA-seq data, used as response matrix, has two dimensions or modes (genes and time points) while the ChIP-seq data, used as predictor variables, includes a third dimension, namely the histone modification type (Figure 3.3).

N-PLS methodology obtains latent components that are linear combinations of the original variables and collects most of the covariance between the response and the predictors. We created a model with three modes (Figure 3.9A, B and C, respectively) and two components that are represented in the X and Y axes of the plots in Figure 3.9. The relationship between the latent components and their interpretation, based on the original modes, is interpreted using the core matrix (Table 3.1). The first element of the N-PLS core revealed that the combination of the second components for each of the three modes (genes, time points and histone modifications) captured most (93.04%) of the gene expression variability. The second most important core element connected component 1 of the first and second modes (genes and time points), with component 2 of the third mode (histone modifications), explaining only 2% of the variability in the data. Figures 3.9A–C show the loading plots for the three modes, that is, the projections of genes, time points and histone modifications onto the new space formed by the N-PLS components.

Table 3.1: Elements of the core matrix sorted by explained variation.

Element	Weight	SS	Expl. Var.
(2,2,2)	0.93	0.87	0.93
(1,1,2)	0.14	0.019	0.02
(1,1,1)	-0.13	0.017	0.018
(1,2,1)	-0.09	0.008	0.008

N-PLS loading plots of mode 1 (genes) and mode 2 (time points) for both RNA-seq and ChIP-seq followed the phases of the YMC, as expected (Figures 3.9A and B). The second component of the second mode separated time-points 5 to 11 from timepoints 1 to 4 and 13 to 15, which corresponded to high and low oxygen consumption stages of the YMC, respectively. The second component of the first mode showed separation of genes at RB phase from genes associated to OX and RC phases; these differences were more pronounced in RNA-seq than in ChIP-seq data, possibly because gene selection and clustering of YMC phases was done on the transcriptomics data.

Regarding histone modifications, the N-PLS loading plot (Figure 3.9C) revealed that the two genomic regions defined to quantify each histone modification (–300 bp to TSS and TSS to +300 bp) were highly correlated, which suggests that a unique region is informative of the chromatin state. Interestingly, the N-PLS results highlight the relevance of H3K9ac and H3K18ac histone modifications on global gene expression regulation, as these two marks had the highest loadings. H3K4me3 was the least important histone modification in global terms, while H3K14ac, H3K56ac and H4K5ac presented intermediate relevance (phases in which these histone modifications peak are displayed in Figure 3.10).

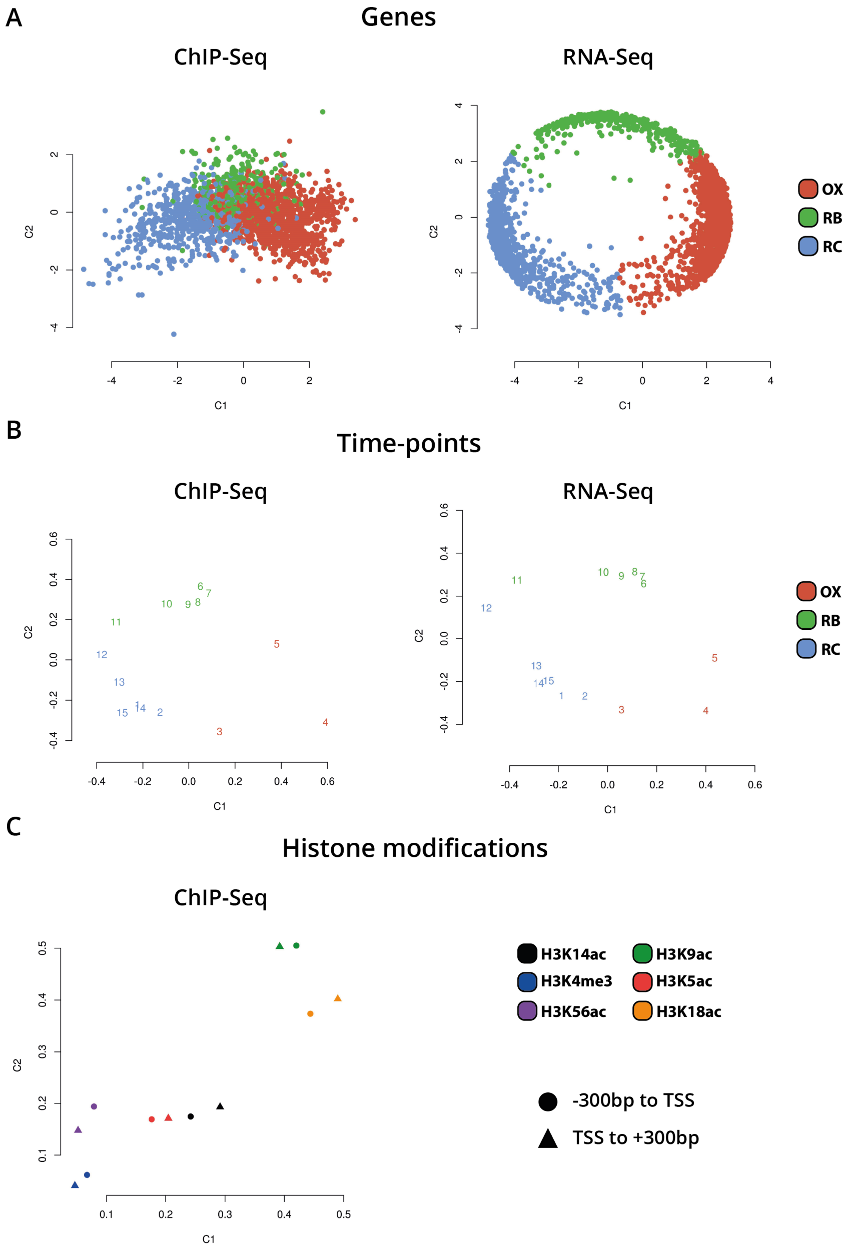


Figure 3.9: N-PLS loading plots for each mode and omic data type. The loading values capture the projection of genes, time points and histone modifications onto the new spaced formed by component 1 (X axis) and 2 (Y axis). Left and right columns display the information for ChIP-seq and RNA-Seq data, respectively. (A) Gene loadings. (B) Time point loadings. (C) Histone modification loadings

3.3.3 MORE models confirmed the relevance of H3K9ac and H3K18ac

In order to unravel the specificity of the histone mark-gene expression regulation, MORE regression models were calculated for each gene using the normalized read count values at the genomic regions defined for each histone mark ChIP-Seq assay as predictor variables. Only significant regulations with positive regression coefficients were used, based on previous studies showing that these histone modifications have a positive regulatory role on transcription activation [10]. MORE results confirmed the relevance of H3K9ac and H3K18ac, as they appear significant in the highest number of gene models: 1042 and 947, respectively. When analyzing the significant regulations per cluster (Figure 3.10), we see that H3K9ac regulates the highest proportion of genes in OX and RB phases (51 and 31%, respectively), followed by H3K18ac in OX (35%) and by H3K56ac in RB (22%). In the RC cluster, H3K18ac is the most prevalent histone modification, regulating 52% of the genes.

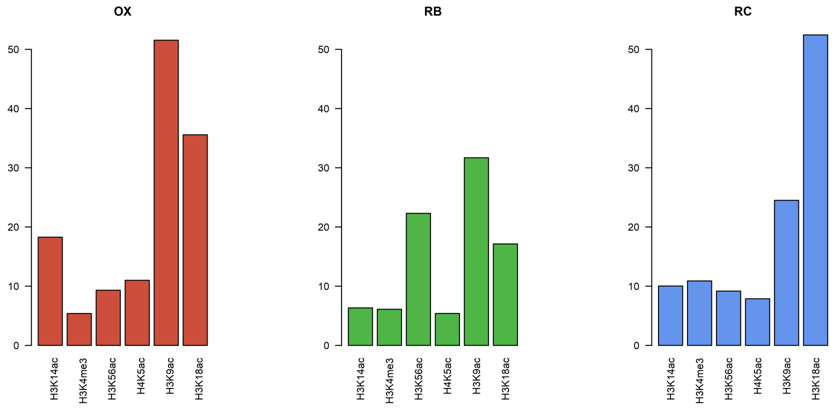


Figure 3.10: Gene expression regulation by histone modifications per YMC phase. Y axes show the percentage of genes in the cluster that was significantly regulated by each histone modification according to MORE models.

In order to understand the impact of histone modifications on the regulation of cellular functions associated with each YMC phase, a functional enrichment analysis of the genes significantly regulated by each histone mark according to

MORE was performed separately for each cluster. The results per phase are described next and are illustrated in Figures 3.11 and 3.12.

3.3.3.1 OX phase

Active histone marks at OX phase regulated translation, ribosomal machinery and nucleotide metabolism genes. Except H4K5ac, every other histone modification was involved in amino-acid metabolism. H3K9ac, H3K18ac, H3K14ac and H3K4me3 were specifically linked with one-carbon metabolism and methylation through the terms *Methyltransferase activity, Asparagine biosynthetic process and Glutathione metabolism*. H3K56ac, H3K18ac and H4K5ac showed overrepresentation of helicase activity genes. H3K18ac and H4K5ac were also enriched in terms related to cell cycle regulation (*Cell division and Mitotic recombination*). Surprisingly, this was not the case for the H3K56ac mark, which was previously reported to be linked to histone deposition in S-phase [116].

3.3.3.2 RB phase

RB phase was characterized by the regulation of sugar metabolism, which is associated to all histone modifications, coupled with a regulation of mitochondrial activity, from which H3K4me3 and H3K56ac were excluded. Genes involved in phospholipid metabolism were enriched in H3K9ac and H3K18ac regulations, while the cell cycle appeared to be controlled by H3K18ac and H3K4me3 marks. H3K9ac, H3K14ac and H4K5ac were significantly enriched in *Biosynthesis of secondary metabolites* functionalities.

3.3.3.3 RC phase

This phase was arguably the most diversified among the histone modification regulatory landscape. The *Tricarboxylic acid* (TCA) cycle was regulated by all modifications through *Oxidative phosphorylation and Electron transport chain*, whereas *Glycolysis* was only enriched at H4K5ac, H3K18ac and H3K14ac. H3K14ac and H3K4me3 appeared to coordinate the regulation of ethanol metabolism genes, while H3K9ac was associated with fatty acid metabolism together with H3K18ac, H3K56ac and H3K14ac. H3K18ac and H3K56ac combined to

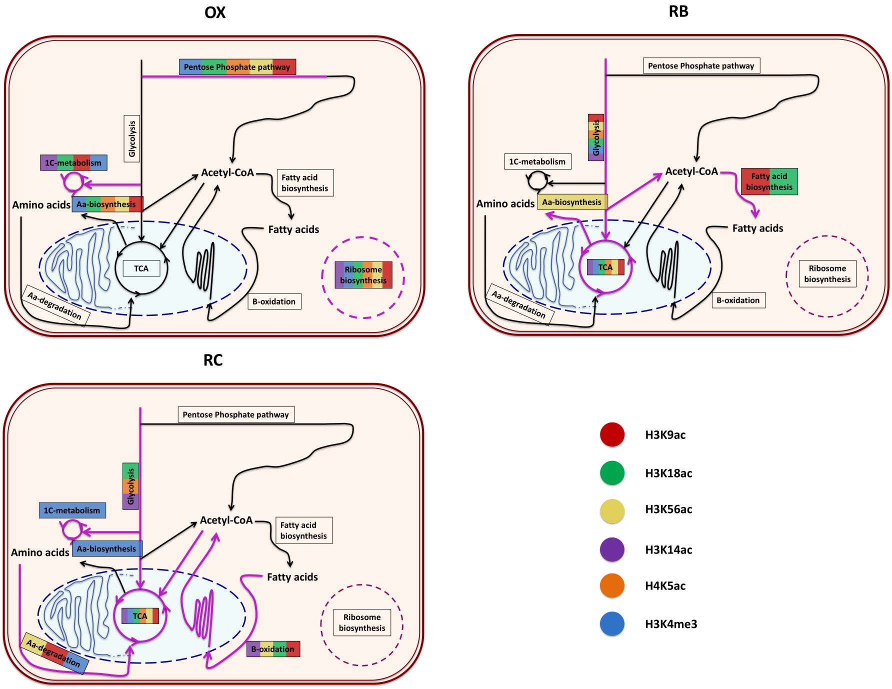


Figure 3.11: Functional enrichment of the genes that were significantly regulated by different histone modifications according to MORE modelling. For each YMC phase, a diagram summarizes the enriched processes found for each color-coded histone modification. The order and the length of the colored lines present in a particular process have no special meaning, other tahn to show that the process was significantly enriched for those histone modifications. The diagrams represent the core metabolic pathways enriched for the different chromatin modifications.

target genes with cell division functionalities, and H3K56ac, H3K9ac and H3K4me3 were related to amino-acid degradation. H3K56ac seemed to be the only mark associated with genes involved in histone acetylation, while H3K4me3 was the only histone modification enriched in one-carbon metabolism genes.

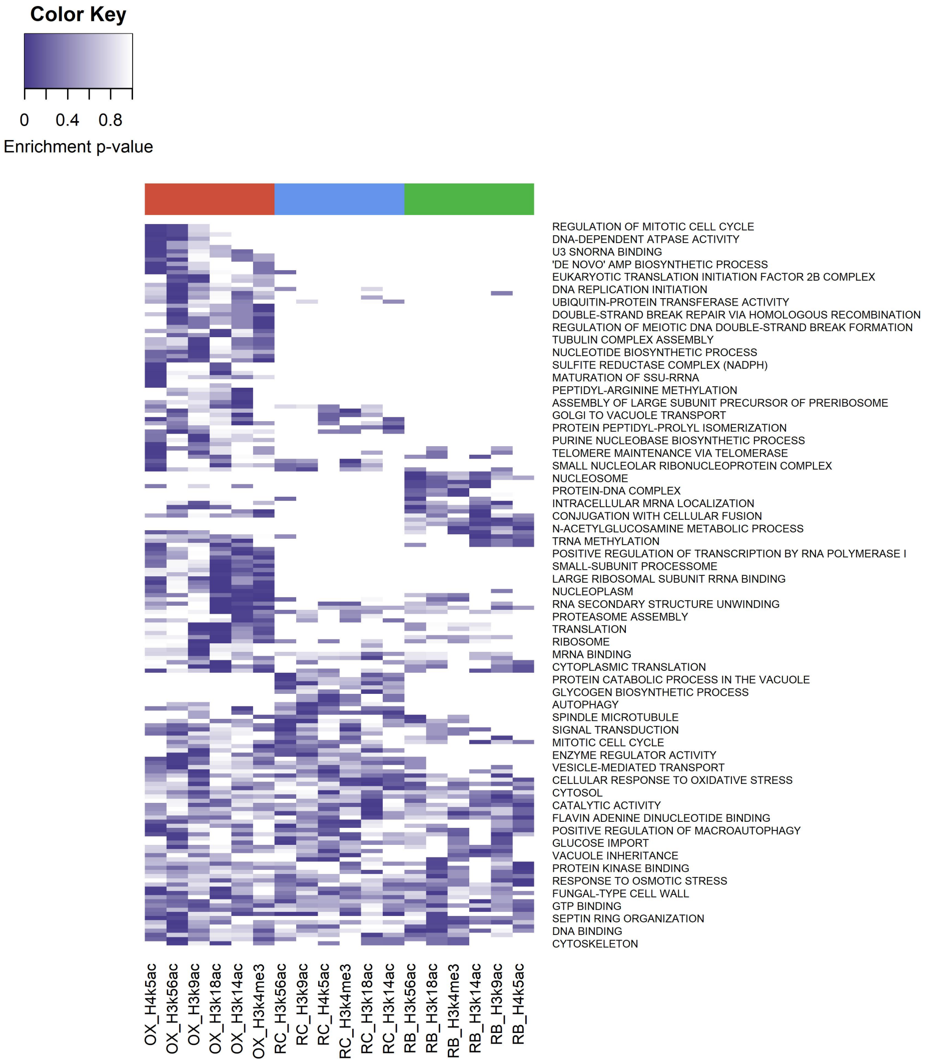


Figure 3.12: Heatmap from the functional enrichment of histone modification target genes from each YMC phase. Rows display the different enriched GO terms, and color intensity represents the p-value of the significancy test. Histone modifications per phase are present in the columns; at the top of each column, three colors represent the three YMC phases (red for OX, green for RB and blue for RC).

MORE identifies most relevant TFs for YMC regulation

TFs are believed to work tightly with histone modifications to regulate gene expression. We used the MORE approach to identify key TFs of the YMC that complement histone mark regulation of gene expression. According to Yeastract database [183], 109 TFs were present among our DEGs. Yeastract also provided the potential target genes for each TF, and this information was used to infer MORE regression models. MORE results indicated that 105 TFs were significantly associated to 2,480 genes. TFs found associated to a significantly higher proportion of genes in the MORE models compared to Yeastract annotations were selected for further analysis, as they represented YMC-enriched regulatory elements. These TFs included Sfp1, Hfi1, Asg1, Ppr1, Ste12, Ylr278c, Cup9, and Dat1 at OX phase; Yhp1 at the RB phase; and Mig2, Pip2, Xbp1, and Cin5 at the RC-phase, making a total of 13 significantly over-represented (FDR adjusted p-value $<$ 0.05) TFs within the MORE models.

Figure 3.13 shows the processes that were biologically regulated by these 13 TFs. In general, TF-regulated genes participated in metabolic sensing, chromatin structure, and cell cycle regulation. Sfp1 appeared to be involved in the transcription of ribosome, nutrient response, G2/M transitions, DNA damage, and histone exchange genes. Hfi1 played a role in translation, amino-acid and nucleotides biosynthesis and nucleotide cleavage via Helicase activity. Pip2 was associated to lipid metabolism, amino-acid and carbohydrate metabolism and the citrate cycle. Mig2 contributed to the regulation of carbohydrate and amino-acid metabolism as well as TCA cycle, and was involved in endonuclease cleavage, mitotic cell cycle and endocytosis. Yhp1-enriched functions were ribosome, helicase activity and cell cycle regulatory control. Xpb1 and Ppr1 were involved in the cell cycle via cyclin regulation, but also in ethanol metabolism and showed regulation of various metabolic pathways, including carbohydrate and amino-acid metabolism. Cin5 was linked to stress response and metabolic response genes; in the context of the YMC, this is mostly related to redox changes and the regulation of glycolysis and amino-acid metabolism. Ste12 regulated

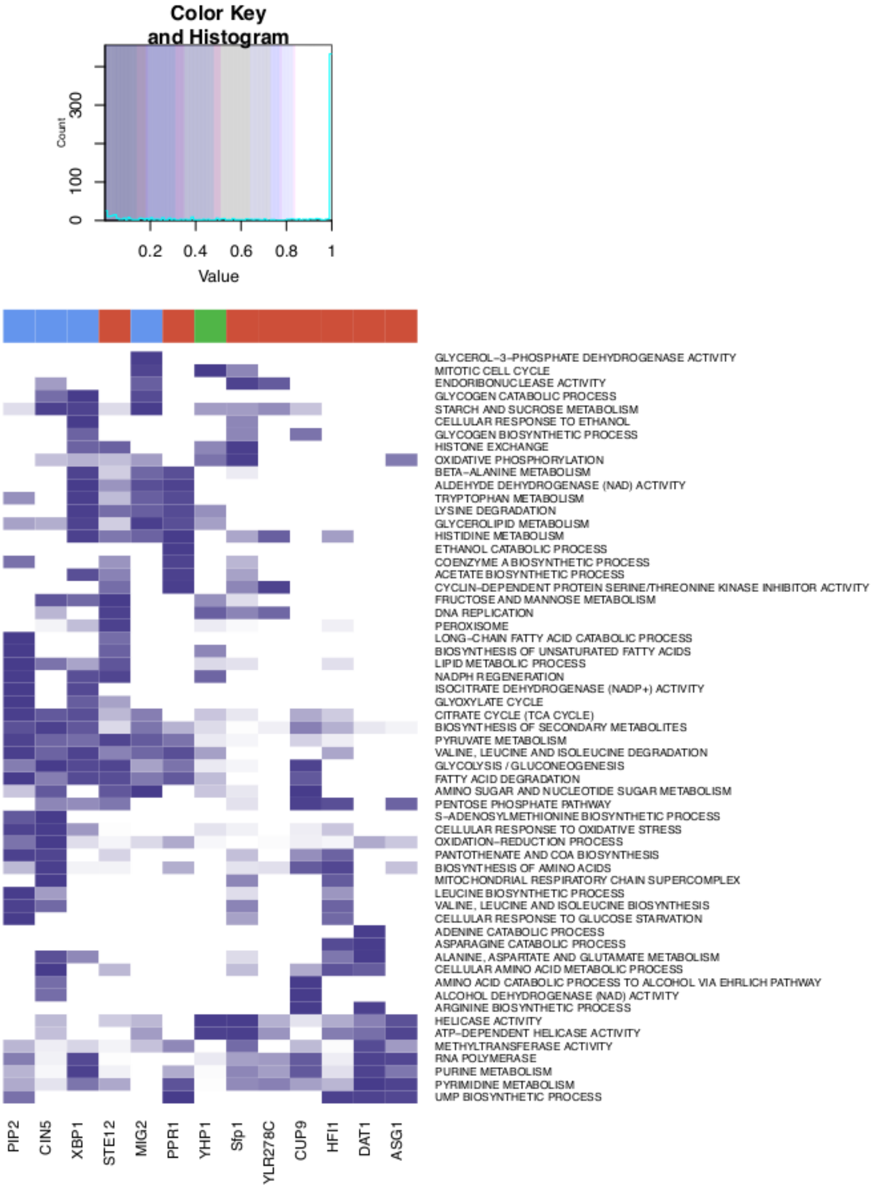


Figure 3.13: Heatmap from the functional enrichments of Transcription Factors (TFs). Rows display the different enriched GO and KEGG terms, and color intensity represent the p-value of the significance test. TFs are present in the columns, at the top of each column three colors represent the three YMC phases (red for OX, green for RB and blue for RC).

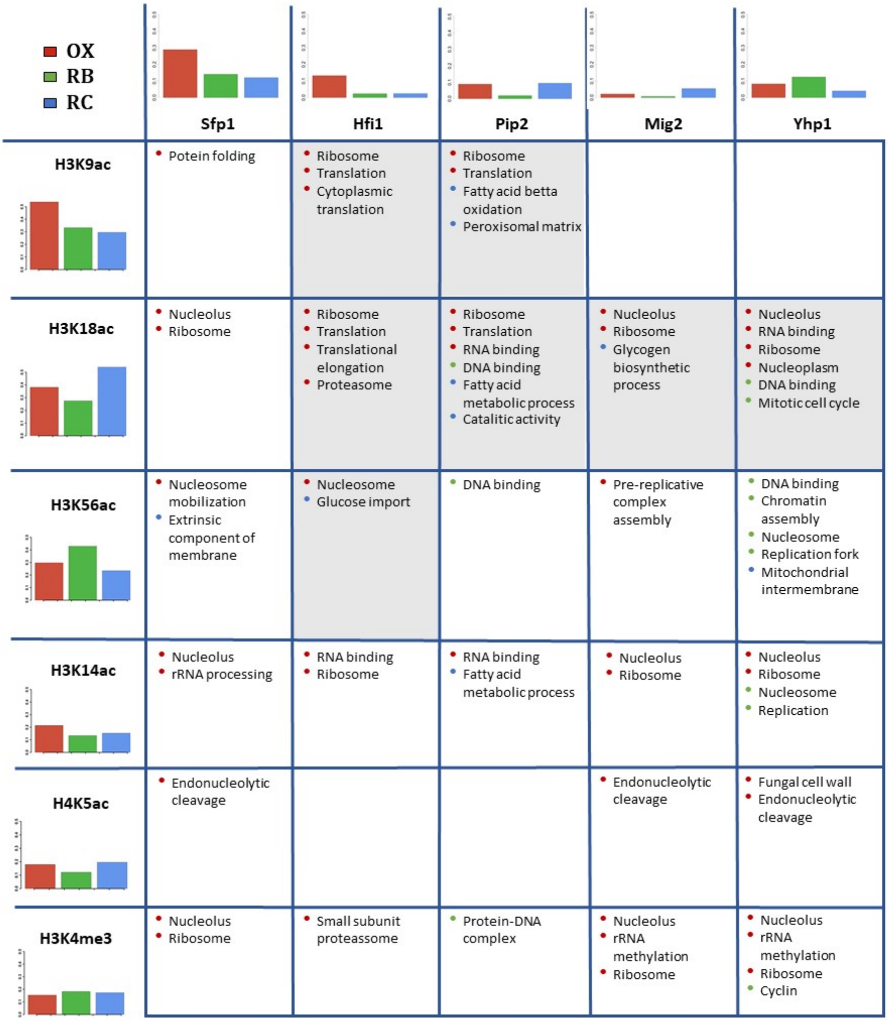
genes in DNA replication and fatty acid metabolism; Cup9 showed enrichment of glycolysis-related functionalities; and Dat1 was involved in metabolic regulations in response to hypoxia, with nucleotide metabolism upregulated, which was also linked to Asg1. Ylr278c showed association with chromatin remodeling functionalities (RSC complex among others), as well as helicase activity and other nucleotide repairing properties.

3.3.4 Gene Expression co-regulation by histone modifications and TFs

After separately studying the role of histone modifications and TFs on gene expression regulation, the results were combined to search for co-regulatory activity between these two types of regulatory elements. First, an independence test was applied to study whether a given pair TF-histone modification was significantly overrepresented in a common set of genes. TFs considered for this analysis were the 13 over-represented factors previously described. Three of all the tested combinations gave significant results: histone modification H3K18ac associated with TFs Hfi1, Pip2, Mig2, Yhp1 and Xbp1; H3K9ac with Pip2 and Hfi1; and H3K56ac with Hfi1 (FDR adjusted p-values $<$ 0.05).

Next, we compared the functional enrichment results obtained independently for histone modifications and TFs (Figure 3.14). For this analysis, we compared the 5 most relevant TFs (FDR adjusted p-value $< 1.1e^{-9}$) with all studied histone modifications. Figure 3.14 includes bar plots indicating the proportion of genes regulated in the YMC phases by each regulator in the table. In general, a higher proportion of regulated genes by histone modifications is expected, as all genes have their corresponding histone modification measurement, but not all the genes have an associated TF in Yeastract database. In OX phase, we found a relative higher number of genes regulated by Sfp1 and Hfi1 together with H3K9ac and H3K14ac. For RB phase, Yhp1 and H3K56ac appeared more frequently, while the number of significantly regulated genes for Mig2 and H3K18ac was higher in RC phase.

Comparative functional analysis revealed functionalities shared between histone modifications and TFs, most of which were consistent with the functions individually attributed to each regulatory element. For instance, H3K9ac, H3K18ac and H3K56ac share OX phase-related functionalities with all the TFs, including *ribosome biogenesis*, *translation* or *helicase activity*, among others. H3K14ac also shared some of these functionalities, but only with TFs Hfi1, Pip2 and Yhp1. RB phase did not show high overlap between histone modifications and TF-associated processes, H3K56ac shared DNA-binding functionalities with Pip2 and Yhp1. Pip2 seemed to be the TF with the highest functional cooperation with histone modifications in RC phase, showing common functionalities with H3K9ac, H3K18ac and H3K14ac, which include *fatty acid metabolic process* and *peroxisomal matrix*.



	Sfp1	Hfi1	Pip2	Mig2	Yhp1
H3K9ac	🔴 Potein folding	🔴 Ribosome 🔴 Translation 🔴 Cytoplasmic translation	🔴 Ribosome 🔴 Translation 🔵 Fatty acid betta oxidation 🔵 Peroxisomal matrix		
H3K18ac	🔴 Nucleolus 🔴 Ribosome	🔴 Ribosome 🔴 Translation 🔴 Translational elongation 🔴 Proteasome	🔴 Ribosome 🔴 Translation 🔴 RNA binding 🟢 DNA binding 🔵 Fatty acid metabolic process 🔵 Catalitic activity	🔴 Nucleolus 🔴 Ribosome 🔵 Glycogen biosynthetic process	🔴 Nucleolus 🔴 RNA binding 🔴 Ribosome 🔴 Nucleoplasm 🟢 DNA binding 🟢 Mitotic cell cycle
H3K56ac	🔴 Nucleosome mobilization 🔵 Extrinsic component of membrane	🔴 Nucleosome 🔵 Glucose import	🟢 DNA binding	🔴 Pre-replicative complex assembly	🟢 DNA binding 🟢 Chromatin assembly 🟢 Nucleosome 🟢 Replication fork 🔵 Mitochondrial intermembrane
H3K14ac	🔴 Nucleolus 🔴 rRNA processing	🔴 RNA binding 🔴 Ribosome	🔴 RNA binding 🔵 Fatty acid metabolic process	🔴 Nucleolus 🔴 Ribosome	🔴 Nucleolus 🔴 Ribosome 🟢 Nucleosome 🟢 Replication
H4K5ac	🔴 Endonucleolytic cleavage			🔴 Endonucleolytic cleavage	🔴 Fungal cell wall 🔴 Endonucleolytic cleavage
H3K4me3	🔴 Nucleolus 🔴 Ribosome	🔴 Small subunit proteassome	🟢 Protein-DNA complex	🔴 Nucleolus 🔴 rRNA methylation 🔴 Ribosome	🔴 Nucleolus 🔴 rRNA methylation 🔴 Ribosome 🟢 Cyclin

Figure 3.14: Co-regulation study for histone modifications and TFs. In rows, the studied histone modifications. In columns, the 5 most relevant TFs. Bar plots show the proportion of genes significantly regulated by each histone modification or TF within each cluster (YMC phase). The table displays the common enriched functional terms for each pair TF-histone modification. The color of the bullets refers to the cluster for which the functional term was enriched. Cells in gray color indicate that the corresponding pair TF-histone modification was co-regulating a significant number of genes.

3.4 Discussion

Temporal profiles of gene expression along the YMC have been widely studied by different authors [97, 167, 187]. However, little is known about the regulatory effect of histone modifications on gene expression in the YMC. The pioneering study of Kuang et al. [102] analyzed chromatin state data and its impact on gene expression in YMC. Although their correlation analysis showed an association between histone modifications and YMC phases, they did not identify which genes were significantly regulated by each histone mark.

In this work, we recovered RNA-seq and ChIP-seq data from the study of Kuang et al. [102] and complemented them with an additional histone modification (H3K18ac). We re-processed ChIP-seq samples to obtain comparable chromatin state measurements and redid the RNA-seq differential expression analysis to refine the clustering of the genes into the three YMC phases. We applied, for the first time in this context, an integrative strategy based on multivariate regression models that allowed us to elucidate interplay of histone modifications and TFs on the modulation of gene expression during the YMC.

Differential gene expression analysis was the starting point of the study, and 2,552 differentially expressed genes obtained with the maSigPro method constituted the set of genes used for downstream analyses. The clustering of these genes into the three YMC phases (Figure 3.6) outperformed previous clustering efforts [102] according to the Silhouette quality indicator 3.8, hence providing a solid landscape to conduct the omics integration analyses.

The multi-omic exploratory approach (Figure 3.9) confirmed that the data was free of outliers and batch effects, and followed the expected distribution of time points and genes in concordance with the YMC phases. The N-PLS results for histone modifications showed that H3K9ac and H3K18ac were the main marks involved in changes of gene expression. Interestingly, these changes were mostly explained by the components that best separated the OX and RC phases from RB. As these differences corresponded to the separation of high and low oxygen consumption, it could be hypothesized that chromatin changes have the highest impact on gene expression through the change in cellular redox state. These results are in line with a YMC division into two phases (HOC and LOC) based on the oxidative state of the cell as proposed in Mellor [126].

While N-PLS allowed for global exploration of the relationship between gene expression and histone modifications, the MORE method provided the framework for dissecting the regulatory program at the gene level by revealing which histone marks and TFs significantly associated with each gene throughout the YMC. MORE models confirmed N-PLS results that pointed to H3K18ac and H3K9ac as the histone modifications with the highest impact on gene expression, since they had significant coefficients in the largest number of gene-specific MORE models (Figure 3.10). Interestingly, while both histone marks showed a similar impact on the regulation of OX phase genes, H3K18ac alone has a stronger regulatory role at the RC phase. We can also highlight here the contribution of H3K56ac to the regulation of the RB cluster.

The functional enrichment analysis revealed that H3K9ac and H3K18ac target genes were mostly involved in ribosome activity and translation in OX phase, in mitochondrial activity and glycolysis in RB phase, and in fatty acid degradation during the RC phase (Figure 3.11). Results for H3K9ac are in agreement with those reported in Kuang et al. [102] (H3K18ac was not included in that study), which were derived from the functional enrichment of genes within each YMC phase. Thus, our findings support these two histone marks as those primarily responsible for driving the clusters' functionalities. Previous studies have pointed to the relevance of acetyl-CoA in coordinating gene expression through protein acetylation levels [198], which peak toward the end of OX phase [21]. This correlates with positive regulation of CHO metabolism as cells enter the RB phase.

We found 13 TFs that are key for the YMC regulation and hypothesize that histone modifications and TFs coordinate to regulate the transcriptional oscillations in the YMC. By looking into the genes that were regulated by each histone modification-TF pair we concluded that Pip2 and Hfi1 combine with H3K9ac and

H3K18ac to drive the OX phase (Figure 3.14). Interestingly, H3K56ac associated with Sfp1, Hfi1, Mig2 and Yhp1 in the regulation nucleosomal functionalities during the OX phase. This putative function in transcriptional control might be related to H3K56ac's role in histone turnover at promoters [116]. Conversely, H3K56ac enrichment in cell division genes during the RC phase might be explained by its role in compaction of DNA into chromatin following DNA replication and repair (Kurdistani and Grunstein, 2003); the role of H3K56ac in replication was previously linked to the RB phase of the YMC [21]. Pip2 is the main regulator of RC phase in coordination with the histone modifications (Figure 3.14), which presumably drive the fatty acid metabolic capabilities of this phase, consistent with the fatty acid response functionalities associated to Pip2 [8, 88]. Overlapping functionalities for H3K9ac and H3K18ac with Pip2 revealed a possible coordination of the marks and the TF to drive fatty acid metabolic processes in RC phase.

Co-regulation analysis revealed that H3K18ac precisely shared a significant number of genes with Pip2, Hfi1, Mig2 and Yhp1. Pip2 triggers the regulation of genes required for beta-oxidation [88], while Hfi1 is required for the acetylation mechanisms of SAGA complex. These results are in agreement with the central concept of the YMC, that explains the link between metabolism and chromatin changes through the expression of lysine acetyl transferases leading to the expression of genes that drive metabolism. This can also be observed at the H3K18ac and Mig2 pair, which share the common function of glycogen biosynthesis (Figure 3.14), an important RC phase pathway. Previous results demonstrating the involvement of Mig2 in the regulation of glucose response [52] supports this hypothesis. Xbp1, Cin5, Ste12, and Cup9 also showed involvement in metabolic response, mostly through fatty acid and glucose metabolism, but they were not associated with any of the histone modifications. Other major cell functions, such as the cell cycle, were associated with TFs (Yhp1, Asg1, Ste12, Cin5) and histone modifications (H3K18ac, H3K56ac, H3K14ac).

Although H3K4me3 was not associated with the regulation of a high number of genes [74], it is worth mentioning that this modification is associated with genes involved in synthesis of amino acids, and, specifically, one carbon metabolism [128]. Histone methylation dynamics have been related to availability of methionine and regulation of one carbon metabolism which, in turn, is the main supplier of S-Adenosyl Methionine (SAM) for protein methylation [128]. H3K4me3 did not show a substantial overlap with any of the TFs, which might suggest a role independent from the selected TFs.

Altogether, the present study combined global and local regression models to unravel the interplay between TFs and histone modifications in the regulation of YMC gene expression. A total of 13 TFs were identified to be particularly active in coordinating YMC progression, while H3K9ac and H3K18ac emerged as the main histone modification drivers acting on the control of fatty acids, amino-acid metabolism, glycolysis and the TCA cycle. More importantly, many coincidences were found between cellular functions enriched at particular TFs and histone mark regulatory programs, suggesting a functional orchestration of these two regulatory layers to drive yeast cell through their metabolic cycle.

Chapter 4

Extracting ATAC-Seq and Metabolomics data from the Yeast Metabolic Cycle to create a multi-layered omics dataset

4.1 Introduction

Gene expression is a tightly regulated process controlled by a complex molecular machinery operating at multiple cellular levels. Traditionally, transcription factor activity has been considered the major regulatory mechanism for gene expression. Recent advances in genomics sequencing technologies, such as ChIP-seq, have unlocked the study of the genome-wide binding of specific TFs to their targeted promoters [141]. Histone modifications have also been known for decades to significantly contribute to gene expression regulation, and similarly, sequencing methods have importantly contributed to establish the epigenomic landscape of the histone marks associated with transcriptional activity [9, 10]. Finally, recent research has identified histone modifications themselves as molecular sensors of cellular metabolic changes and points to a flux of molecular information that connects metabolic changes to the regulation of gene expression [198].

The Yeast Metabolic Cycle (YMC) is a powerful model system for the study of celular regulation. In this system, yeast cells are synchronized and cultured in highly controlled fermenter conditions, during which O_2 consumption [126] is monitored. During YMC, metabolites cycle with similar oscillatory patterns as those seen in gene expression and histone modifications. The influence that metabolites have in YMC progression has been mainly analyzed within the framework of acetyl-CoA metabolism [21, 188]. The addition of acetate to the fermenter causes the yeast cells to enter a High Oxygen Consumption (HOC) state and triggers a peak in histone acetylation, which points to metabolic regulation of gene expression in the YMC [21].

The YMC is a unique system in which to address the development of multilayered systems biology models of transcriptional regulation. The exceptionally high reproducibility of this experimental system facilitates the generation of different molecular profiling datasets that can be combined by matching samples along the time course using O_2 consumption level at each sampling times. This point-by-point sample matching is critical for the application of correlation-based data analysis strategies [181].

Thanks to these experimental possibilities, a wealth of molecular matched data exist today for the YMC, which includes multiple histone marks and gene expression described in Chapter 3. However, currently, there are no complete metabolomics datasets that cover the totality of the cycle with high resolution. Moreover, the chromatin structure, and its dynamic changes across cycle, have not been yet assessed, which represents a limitation in the study of the metabolic control of chromatin accessibility and transcriptional activation.

Chromatin structure can be measured using ATAC-seq, a molecular technique that detects the naked parts of the DNA that are free of histones and are therefore accessible to transcription factor binding [19]. In ATAC-seq, cell nuclei are isolated and subjected to transposase activity, which digests DNA that is not occupied by proteins. Although this technique has previously been used in multiple cell cultures, including yeast and mammalian systems among others, it has never been applied to the Yeast Metabolic Cycle, where oscillations in the thickness of the cell wall imposes particular experimental challenges.

In this chapter, we address the generation of high-quality metabolomics and ATAC-seq data for the YMC with sufficient resolution for efficient integration with current RNA-seq and ChiP-seq datasets. We describe the protocols developed to obtain metabolomics and ATAC-seq measurements from the YMC, as well as the data processing and analyses resulting in the identification of metabolites that cycle during the YMC and of TFs that bind to open chromatin regions. We also incorporate a recently published [51] NET-seq dataset that measures nascent transcripts with polymerase binding sites as an extra information layer on active transcription.

4.2 Methods

4.2.1 Fermenter and data extraction

The yeast strain used for this study was CENPK113-7D rpb3-FLAG. Cells were cultured in a BioFlo320 fermenter (2L vessel, 1.1 L culture volume – New Brunswick), cultures were grown in YMC-YE media (pH 3.5 - ammonium sulphate 5 g/L, potassium dihydrogen monophosphate 2 g/l, magnesium sulphate 0.5 g/L, calcium chloride 0.1 g/L, yeast extract 1 g/L (Difco), glucose 10 g/L, sulphuric acid 0.035 %, antifoam-204 0.05 % (Sigma Aldrich), iron sulphate 20 mg/L, zinc sulphate 10 mg/L, manganese chloride 1 mg/L, copper sulphate 10 mg/L). Fermenter runs had an aeration rate on 1 L/min and agitation rate of 1000 rpm. Run temperature was set to 30ºC and pH was constant at 3.5 through the addition of 0.25 NaOH. 10 mL of starter culture were used to start each fermenter run. Cultures were grown overnight to saturation at 30ºC; after saturation, cells were starved for 6 h, after which, continuous culture was maintained at a flow rate of 1.5 mL/min. The cycle was allowed to stabilize for at least 24 hours between cycling and sampling.

4.2.2 Metabolomics experiments

For metabolomics sampling, 9 tubes yeast cells with with $20x10^6$ per sample were harvested from the BioFlo fermenter. Cells were pelleted, removed from the supernatant and snap-frozen with liquid nitrogen. A total of 107 samples were extracted and evenly distributed in 5 cycles of the YMC. This resulted in 21 timepoints distributed in 5 cycles with 2 extra samples (Figure 4.1A).

Cell pellets were lyophilized to dryness overnight. The lyophilized pellets were homogenized in 50:50 acetonitrile/water with 0.3% formic acid for OAs and AAs, 5% TCA for acetyl and malonyl CoA, 0.5M PCA for oxidized nucleotides and 50:50 methanol/NaOH for reduced nucleotides using the Precellys bead-based homogenizing system. Precellys tough micro-organism lysing kit (VK-05) was used to burst the yeast cells. An aliquot of homogenate for the required assay was aliquoted and immediately stored at -80ºC.

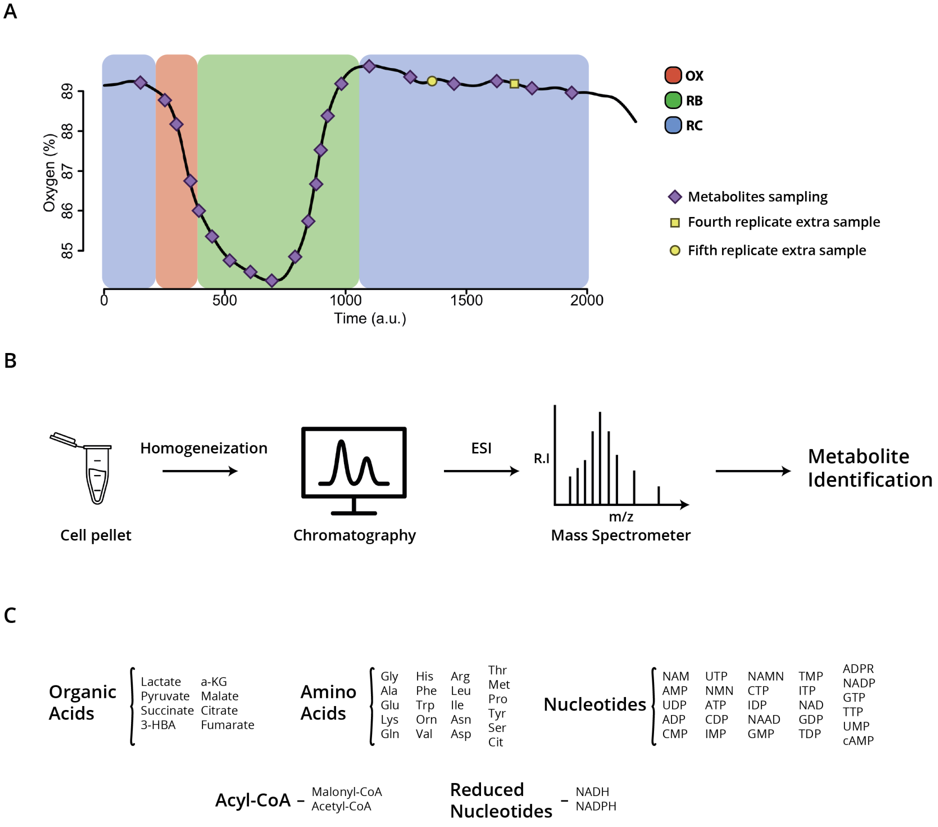


Figure 4.1: (A) Sampling timepoints used to extract the metabolomics dataset. (B) Mass spectrometry analysis pipeline that leads to metabolite identification. (C) List of metabolites measured in the present study.

Oxidized nucleotides were separated on a Thermo Scientific Hypercarb column (3 µm x 50 mm x 2.1 mm) using a Thermo Dionex Ultimate 3000. Reduced nucleotides were separated on a Waters HSST3 1.8 µm x 2.1 mm x 150 mm column. Organic acids were separated on Waters Acquity UPLC BEH C18 2.1 mm x 100 mm, 1.7 µm column. Amino acids were separated on a Waters AccQTag 2.1 x 100 mm, 1.7 µm column. CoAs were separated by a Waters Acquity UPLC BEH C18 1.7 µm x 2.1 mm x 50 mm column. A total of 63 metabolites were separated via chromatography and quantified using a Thermo Scientific Quantiva Triple Quadrupole mass spectrometer (Figure 1B). The mass spectrometer was operated in positive ion mode using electrospray ionization. The raw data was processed using Xcalibur 3.0.

4.2.3 Metabolomics data analysis

A total of 21 matching time-points in each of the 5 measured cycles were selected for downstream analysis. Metabolomics data were first corrected for protein abundance, and then were normalized using a median normalization. The most extreme value of each metabolite was discarded, and the resulting matrix was centered.

Metabolites oscillation were modeled with the R package maSigPro [26, 138] using a polynomial degree of 3. Differentially abundant features were called from significant models (FDR adjusted p-value $<$ 0.05) with a minimum R^2 value. Two thresholds for R^2 were evaluated: 0.4 and 0.6. The abundance profiles of significant metabolites were clustered using K-means. We evaluated clustering in either 2 or 3 groups to assess the current models of YMC oscillations.

4.2.4 ATAC-seq experiments

Yeast cells for ATAC-seq were evenly sampled in 18 timepoints across the cycle in two batches (Figure 4.2A). Cells were fixed in 1% formaldehyde for 5 min at room temperature, glycine was then added to a final concentration of 125mM and incubated for 5 min. An equivalent of 30 million cells were washed with water, resuspended in Zymolyase preincubation buffer (250µL 14.5M β-ME, 27.8µL 0.5M EDTA pH 8 and complete to 5mL with Sorbitol 1M) and kept at 30º for 30 min. Cells were washed in sorbitol, and the pellet was resuspended in 500µL of Zymolyase buffer (3.5µL 14.5M of B-ME, complete to 10mL with Sorbitol 1M, resuspend Zymolyase to a final concentration of 10mg/mL) and incubated 30º for 12 min to digest yeast cell wall. Cells were gently washed twice in sorbitol and twice in PBS to eliminate β-ME remains and then were resuspended in lysis buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630) and centrifuged at 10,000 rpm for 10 min (Figure 4.2B).

The supernatant was discarded, and the pellet included isolated nuclei that were resuspended in 50µL of transposition mix and incubated at 37º of 30 min. When the transposition reaction finished, 10µL of STOP buffer (85% H2O, 5% SDS, 10% 0.5M EDTA) was added to the tubes, DNA was de-crosslinked using

7µL of NaCl 5M and incubated for 3h at 65º. Then, 1µL of proteinase K was added and incubated at 65º overnight.

DNA was purified using a Qiagen MinElute PCR purification Kit, and transposed DNA was eluted in 11µL of elution buffer. DNA was PCR amplified using Nextera DNA Library Preparation Kit and Nextera Index Kit and purified using AmpureXP cleanup for a final insert size of 150-180 bp. Quality of the library size was assessed using Qubit, and libraries were sequenced with Illumina NextSeq 500 with a NextSeq 500/550 High Output Kit v2.5 (75 Cycles).

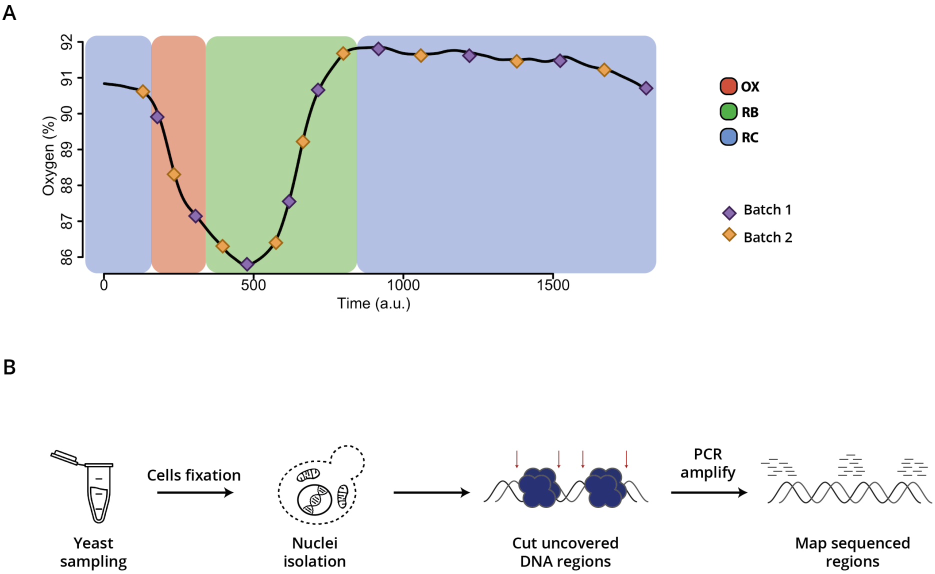


Figure 4.2: (A) YMC ATAC-seq sampling, which was distributed in two batches; sampling was evenly distributed among the three functional phases of the cycle. (B) ATAC-seq protocol, which entails four main processes after yeast sampling: nuclei isolation; transpose-mediated cutting of open DNA; isolation, amplification and sequencing of the cut regions; and mapping of the sequences to the reference genome.

4.2.5 ATAC-seq data analysis

4.2.5.1 Pre-processing

ATAC-seq data quality was assessed with Fastqc [2] and adapter and quality trimmed with Trimmomatic [14] using a minimum quality of 30, a sliding window of 5 and a minimum length of 28 bp. Data was aligned to the Ensembl *S. cerevisiae* reference genome (release 91) using Bowtie [105] mapper. Mapped reads were used for peak calling with MACS software [209], and peaks were subjected to two parallel analyses: detection of transcription factors that bind to the detected peaks and extraction of a count matrix for determination of peak oscillations.

4.2.5.2 Peak calling and quantification

Peaks identified with MACS across samples were merged with BEDtools and quantified using feature counts [118]. The peak quantification matrix was batch corrected with Combat [113], TMM normalized and centered.

Differential peaks were detected by modeling ATAC oscillations with the R package maSigPro [26, 138] using a polynomial degree of 3. Differentially expressed features were called from significant models (FDR adjusted p-value $<$ 0.05) with a minimum R^2 value of 0.4 and 0.6. Differential features were scaled, centered and clustered with k-means into 2 and 3 clusters.

4.2.5.3 Motif search and TF identification

TF binding sites (TFBS) in called peaks were identified using Wellington [145] and sequences were searched for known TFBS with the FIMO software [62], using JASPAR database as the source of known motifs (JASPAR2018_CORE_fungi non_redundant_pfms_meme). The TFs that bound to peak regions were associated with target genes by proximity of the peak to their promoters, and association was done using RGmatch software [56].

4.2.6 NET-seq data analysis

NET-seq count data was obtained from [51]. We applied TMM normalization, log transformation and centering. Differential expression of NET-Seq data was calculated with maSigPro using a polynomial degree of 3, and significant features were called with a minimum R^2 value of 0.6 and p-value $<$ 0.05.

Differential features were clustered with k-means (k = 3) to mimic the three YMC phases.

4.2.7 Tucker3

Consider a three-dimensional $\underline{\mathbf{X}}$ data matrix (*nxmxp*) where *n* are the genes, *m* are the transcription factors and *p* are the timepoints. The matrix represented the binary information of whether the footprint of transcription factor *m* was detected in a peak laying in the promoter of gene *n* at timepoint *p*. The Tucker3 models are considered to be a three-way extension of PCA, allowing to decompose these cubic arrangements in its modes, as shown in equation 4.1.

$$\mathbf{Y} = \mathbf{A}\underline{\mathbf{G}}(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{R} \tag{4.1}$$

Where $\mathbf{A}$ (*nxd*) is the loading matrix corresponding to the first mode, $\underline{\mathbf{G}}$ (*dxexf*) is the core matrix that explains the relationships between the different modes, $\mathbf{C}$ (*mxe*) and $\mathbf{B}$ (*pxf*) are the other two loading matrices and $\mathbf{R}$ represents the residuals. The model parameters are estimated by minimizing the residuals' sum of squares $\sum r_{nmp}^2$.

4.3 Results

4.3.1 YMC sampling of ATAC-seq and metabolomics data

In order to extract ATAC and metabolomics data from the Yeast Metabolic Cycle, yeast cells were cultured in a fermenter with continuous nutrient-limiting conditions, where medium flux, aeriation rate, pH and temperature (among other conditions) were monitored. Controlled conditions make the YMC highly reproducible, and the dissolved oxygen levels allow for cycle progress to be tracked. The YMcharacC is terized by three functional phases of gene expression, which oscillate in High Oxygen Consumption (HOC) and Low Oxygen Consumption (LOC) environments.

We sampled the YMC to obtain metabolomics and ATAC-seq datasets covering all functional phases of the cycle. The YMC was sampled in quintuplicate for metabolomics. This allowed us to obtain robust estimates of metabolite levels for this highly variable data type (Figure 4.1). ATAC-seq data was sampled in two batches (Figure 4.2). We used oxygen profiles to align metabolomics and ATAC-seq samples with existing YMC datasets.

4.3.2 Metabolomics data exploration

4.3.2.1 Exploratory analysis

Metabolomics data preprocessing included correction by protein content, median normalization, elimination of batch effect, centering and scaling. PCA of corrected values revealed that the metabolomics samples were generally grouped according to their YMC phase (Figure 4.3A) and no batch effect was observed (Figure 4.3B). Moreover, the PCA recapitulated the order of extraction of the metabolomic samples across the cycle, suggesting that the experiment also captured the temporal variability of the YMC (Figure 4.3C). The first and second components captured 28.4% and 15.2% of the metabolic variability, respectively.

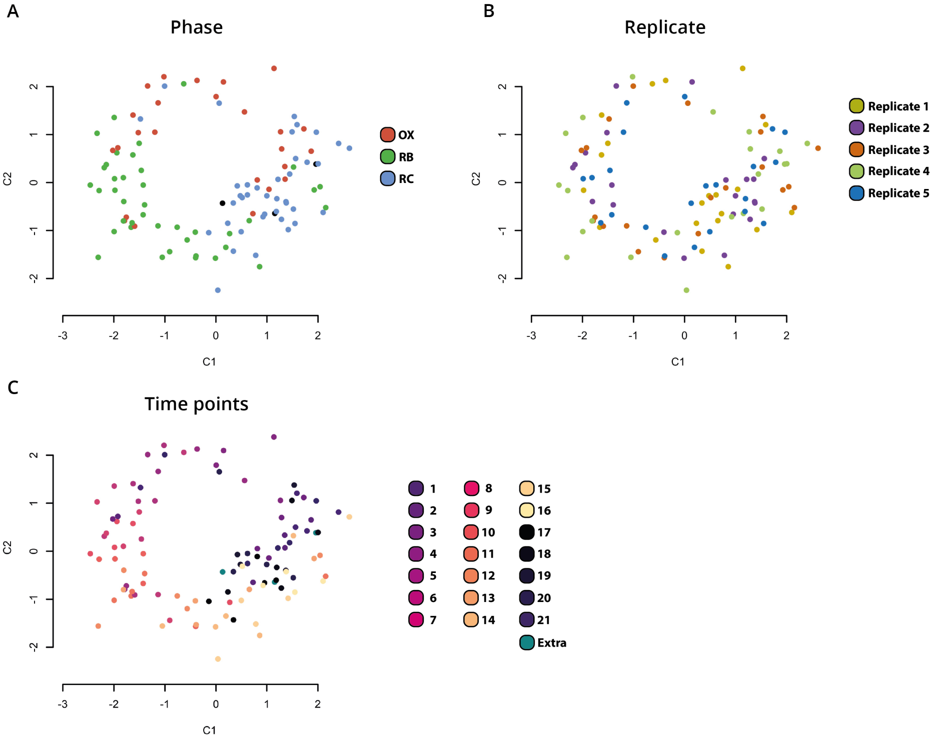


Figure 4.3: PCA of the five metabolomics replicates (total of 107 samples) colored by their YMC phase (A), the replicate they belong to (B) or the timepoint (C).

4.3.2.2 Metabolic profiles

We analyzed variability in the measurement of metabolite levels. For this, we first matched time points across batches based on the oxygen level of each sample. Then, we obtained boxplots for each metabolite at each time point. We observed that, in many cases, at least one clear outlier was present (Figure 4.4). This was not surprising, since metabolomics measurements are known to have a high associated signal-to-noise ratio. To ensure that this variability would not jeopardize downstream analysis, and to take advantage of our level of replication, we applied a procedure where we excluded the most extreme value at each time point to calculate a robust mean. This resulted in metabolite profiles that followed oscillatory pattern across the cycle and showed accumulation at a specific YMC phase (Figure 4.5). ATP accumulated in the early OX phase,

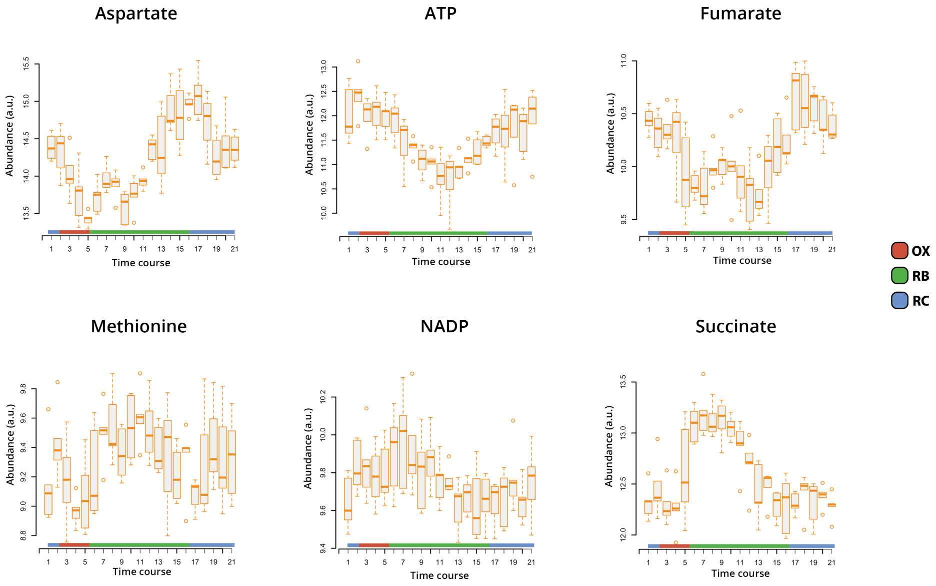


Figure 4.4: Boxplot of six metabolites, showing the mean abundance across the 5 sampling events. The boxplot reveals outliers at each sampling point.

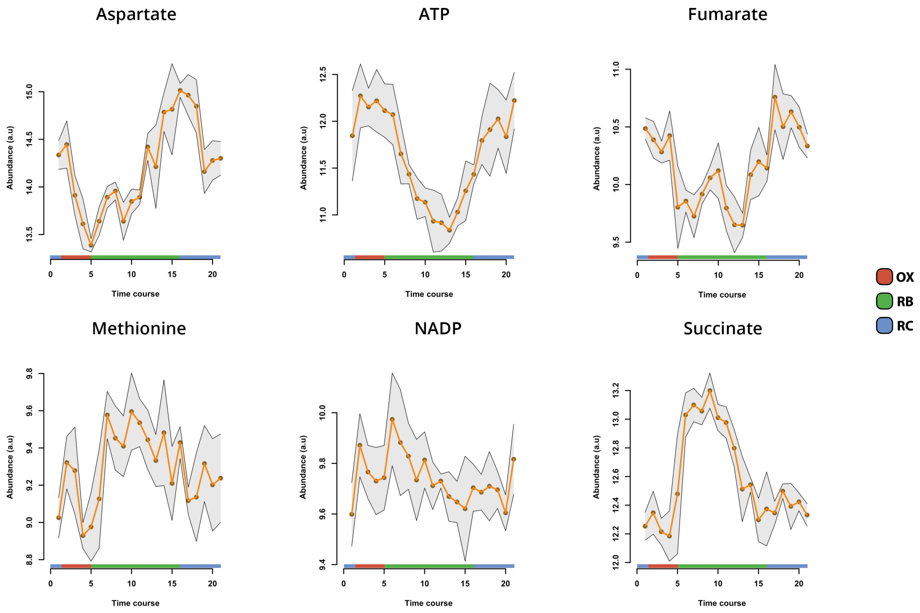


Figure 4.5: Profile of six metabolites after removal of the measurement at each timepoint that deviated the most from the mean. Mean abundance is shown in orange and the mean +- the standard deviation is shown in grey.

while succinate accumulated in early RB. Other metabolites, such as NADP, did not follow a clear oscillation, and the pattern tended to be flat. PCA of this robust mean data showed that the full metabolomics dataset recapitulated the dynamics of the YMC and its phases (Figure 4.6A and Figure 4.6B). These normalized, robust mean data were used in follow-up analyses.

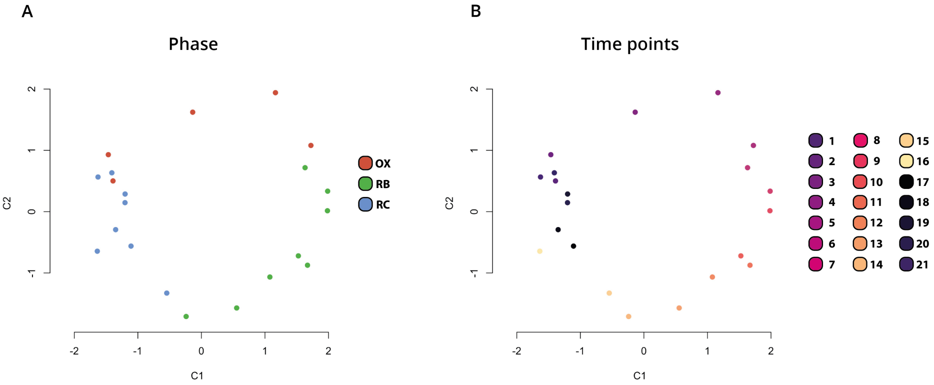


Figure 4.6: Principal Component Analysis of the metabolomics matrix, which was poduced using the mean measurement of abundance of the five replicates at each timepoint and deleting the measurement that deviated the most from the mean. Samples in the two PCAs are colored by the YMC phase they belong to (A) and the timepoint when they were extracted (B).

4.3.2.3 Selection of maSigPro parameters for differential abundance

Metabolite oscillations were modeled using the maSigPro R package [26, 138], which emplys a two-step regression strategy to select metabolites with significant temporal changes. Genes and histone modifications extracted from the YMC and analyzed in the previous chapter were also modeled in maSigPro, and differentially abundant features were obtained when an R^2 of 0.6 was specified.

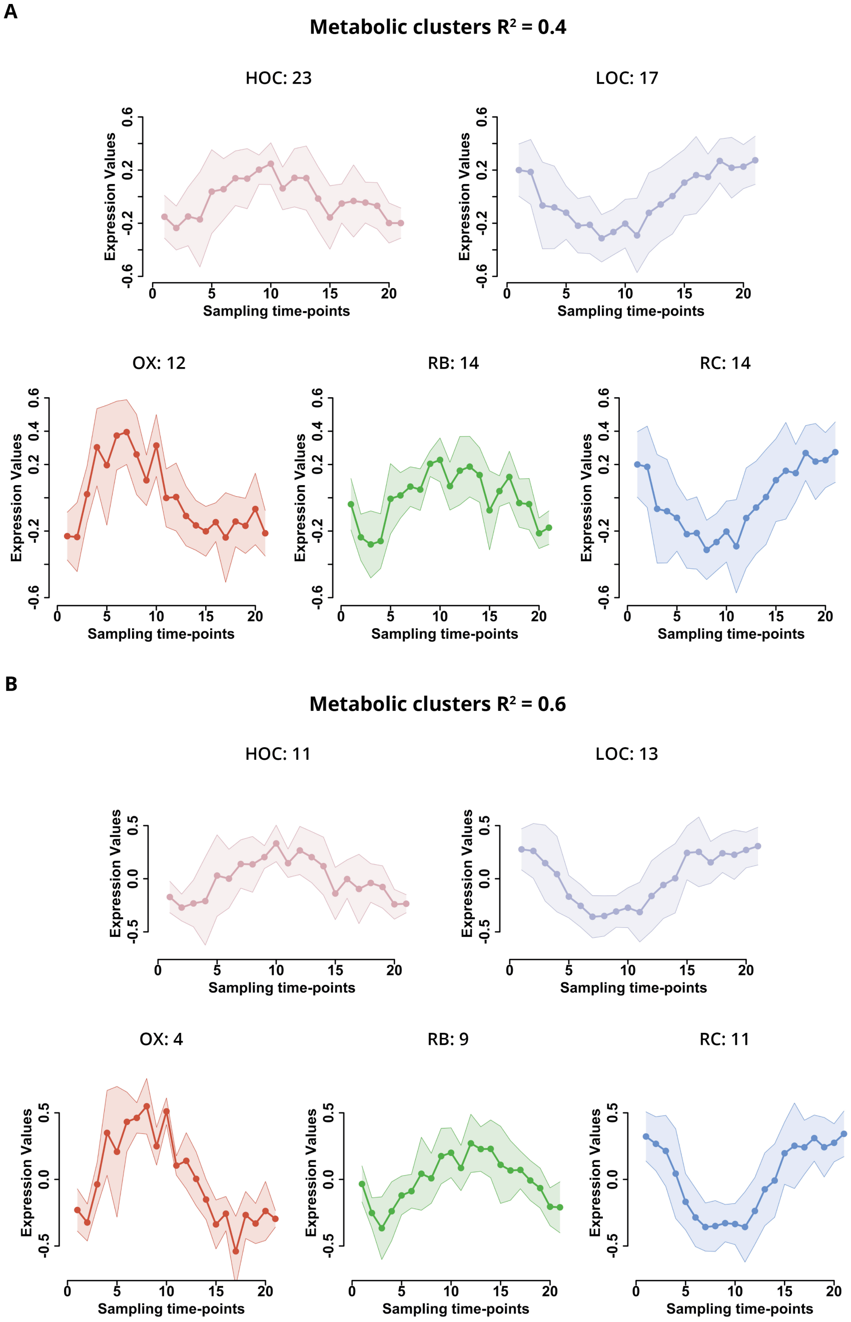


Figure 4.7: Differential expression and clustering of metabolites detected in the YMC. The DE expressed metabolites were obtained using an R^2 of 0.4 (A) and 0.6 (B). The two procedures were compared, revealing that 0.4 could be a more reliable threshold given the signal/noise ratio of these measurements.

For the determination of metabolites with differential abundance, the same R^2 was applied, which resulted in selection of 24 out of 62 metabolites that were differentially expressed over time (Figure 4.7B). Since metabolomics measurements generally have a higher fluctuation and background signal, we also tested an R^2 of 0.4, which resulted in 40 significant metabolites (Figure 4.7A). We performed a clustering analysis of the differentially abundant metabolites using both R^2 measurements independently to aid in the decision of which R^2 was more successful in modelling the YMC oscillations.

The Yeast Metabolic Cycle has been traditionally studied within three functional phases defined by gene expression oscillations (OX, RB and RC). However, recent research has pointed towards a two-phase activity that reflects the High Oxygen Consumption (HOC) and Low Oxygen Consumption (LOC) states that the cycle traverses. We have thus clustered the differentially expressed metabolites in two and three clusters using both sets of metabolites (R^2 of 0.4 and 0.6) to determine which clustering better reflects the metabolic oscillations in the YMC.

Two-phase clustering groups contain 23 and 17 metabolites in HOC and LOC phases, respectively, using an R^2 of 0.4 (Figure 4.7A, top), and 11 and 13 using an R^2 of 0.6 (Figure 4.7B, top). Three-phase clustering groups contained 12, 14 and 14 metabolites in OX, RB and RC phases, respectively, using an R^2 of 0.4 (Figure 4.7A, bottom) and 4, 9 and 11 using an R^2 of 0.6 (Figure 4.7B, bottom). The clustering results are highly similar using an R^2 of 0.4 and 0.6; the main difference is in the number of metabolites.

The comparison between two and three phases is less clear, the main difference being that the peak in OX phase seems to be masked and distributed in the two phases (mainly in HOC). We decided to continue our analysis using the metabolites obtained with an R^2 of 0.4 and clustering to three phases, since it seems to better capture the oscillatory behavior of the cycle.

4.3.2.4 Clusters composition

Each of the three clusters obtained using an R^2 of 0.4 contained metabolites that correspond to clearly differentiated biological processes (Table 4.1). OX phase contains mainly amino acids and citrate; RB phase includes NMPs, NAM metabolites and most TCA intermediates; while RC phase metabolites include NTPs and acyl-CoAs. In the next chapter, metabolites will be statistically integrated with RNA-seq to aid in the functional interpretation of the metabolites' accumulation in the context of the YMC.

Table 4.1: Metabolites that belong to each of the clusters at the two Q2 thresholds used in maSigPro analysis. Here we show the metabolites that belong to the three-cluster and the two-cluster distributions.

Phase	Metabolites	
	R^2 = 0.4	R^2 = 0.6
OX	Lactate, Succinate, Citrulline, Glycine, Leucine, Lysine, Methionine, Ornithine, Proline, NADP, AMP, UMP	Citrulline, Glycine, Lysine, Ornithine
RB	Citrate, Aspartate, Phenylalanine, NAM, NAMN, NAAD, cAMP, ADPR, GMP, GDP, CMP, TMP, IMP, IDP	Citrate, Aspartate, NAM, NAMN, ADPR, GMP, GDP, CMP, TMP
RC	3-HBA, Malate, Fumarate, Isoleucine, Tyrosine, Valine, NAD, ATP, GTP, CDP, CTP, UTP, acetyl-CoA, Malonyl-CoA	3-HBA, Malate, Fumarate, Isoleucine, Valine, ATP, GTP, CTP, UTP, acetyl-CoA, Malonyl-CoA
HOC	Lactate, Succinate, Citrate, Citrulline, Glycine, Leucine, Lysine, Methionine, Ornithine, Proline, NAM, NAMN, NAAD, NADP, AMP, cAMP, ADPR, GMP, CMP, UMP, TMP, IMP, IDP	Citrate, Citrulline, Glycine, Ornithine, NAM, NAMN, ADPR, GMP, CMP, TMP
LOC	3-HBA, Malate, Fumarate, Aspartate, Isoleucine, Phenylalanine, Tyrosine, Valine, NAD, ATP, GDP, GTP, CDP, CTP, UTP, acetyl-CoA, Malonyl-CoA	3-HBA, Malate, Fumarate, Aspartate, Isoleucine, Valine, ATP, GDP, GTP, CTP, UTP, Acetyl-CoA, Malonyl-CoA

4.3.3 Analysis of ATAC-seq data

4.3.3.1 Quality assessment

Quality of samples was first assessed using fastQC software, which revealed that samples 10 and 11 had a low coverage and were highly enriched in adapter sequences, suggesting a low amount of purified DNA at the amplification step. Mapping Quality also indicated poor performance for samples 10 and 11 and revealed that 8 and 13 have a low percentage of mapped reads. Low mapping of total reads generally correlates with low number of detected peaks (Figure 4.8A), thereby limiting the capacity of these samples for capturing the YMC chromatin dynamics. These four samples were omitted from further analyses.

As expected, the number of detected peaks correlated with the total number of mapped reads of the sample (Figure 4.8C).

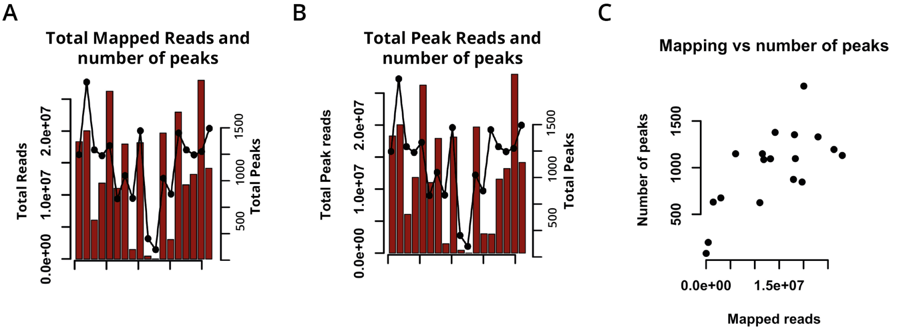


Figure 4.8: (A) Number of mapped reads and detected peaks per sample. (B) Number of reads that mapped to detected peaks and number of detected peaks per sample. (C) Scatter plot of the number of mapped reads and the number of peaks, which shows a correlation between the number of peaks and the number of mapped reads

4.3.3.2 Exploratory analysis

ATAC-seq peak counts were normalized by TMM, centered and scaled to explore variability in the peak quantification across the cycle. Principal Component Analysis (PCA) revealed that sample 12 might be an outlier, and therefore this sample was also excluded from the dataset. After removal of samples 8, 10, 11, 12 and 13, the PCA showed that samples group by their corresponding phase, suggesting that ATAC-seq data captured the variability of the YMC (Figure 4.9). RC phase samples grouped better than those from OX and RB. The second component separated HOC phase samples from LOC phase samples and, overall, samples followed the sampling time course. The first component explains 49.5% of the variance, while the second component explains 15.6%.

4.3.3.3 Analysis of differential peak selection by R^2 of 0.4 and 0.6

ATAC samples were normalized and centered as stated in the methods, and samples 8, 10, 11, 12 and 13 were removed from the dataset due to inconsistencies in the number of sequenced reads, the number of mapped reads (Figure 4.8) or due to the appearance of outliers in the Principal Component Analysis

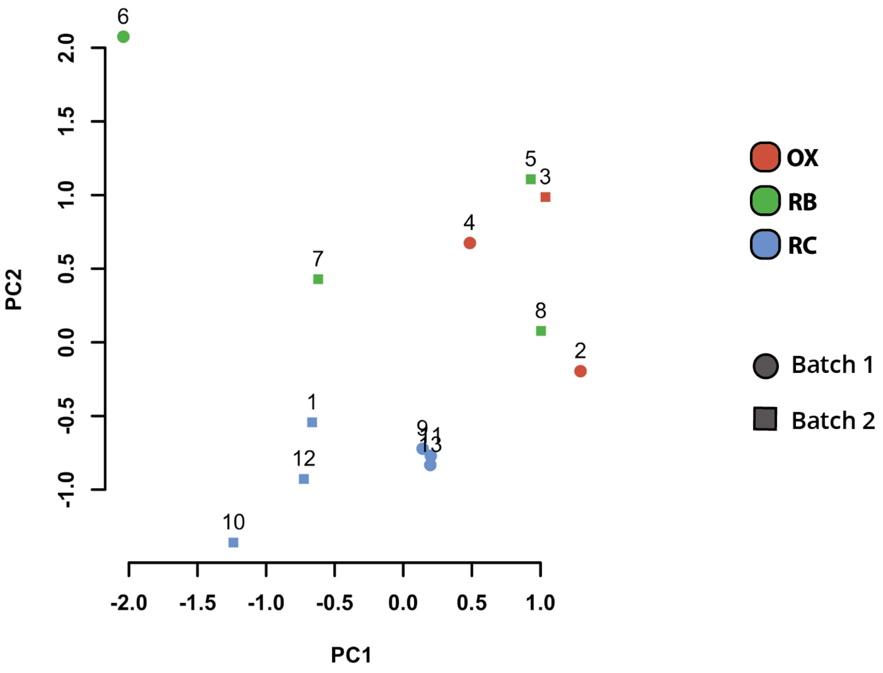


Figure 4.9: Principal Component Analysis (PCA) on the ATAC-seq quantification matrix allowed for exploration of the data variability. PCA shows that samples not only group by phase, but also that there seems to be no batch effect in at least the two components analyzed.

(Figure 4.2). We used maSigPro to model the changes in peak size across the YMC and identify regions with significant changes, which would indicate chromatin reorganization in those positions of the genome.

For the determination of peaks that show significant changes across the cycle, we used R^2 cutoff values of 0.4 and 0.6. A total of 334 peaks had a significant model when using an R^2 of 0.4 or higher (Figure 4.10A), while 208 were identified when using an R^2 of 0.6 or higher (Figure 4.10B). Both sets of genes were then clustered into two or three groups to explore consistency in temporal patterns of change.

4.3.3.4 Peaks clustering

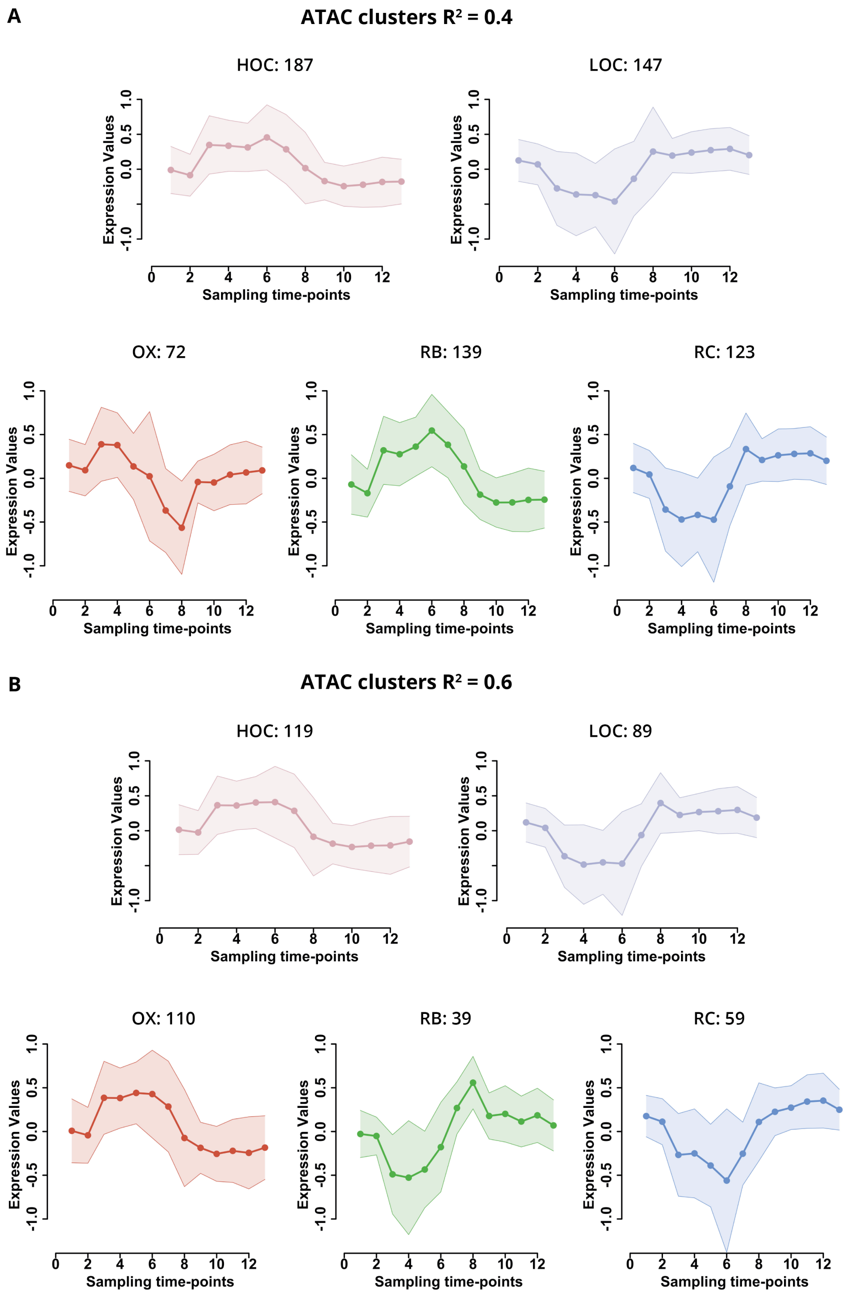
Peaks were clustered in two and three clusters mirroring the metabolite analysis. The two clusters recapitulated the HOC and LOC phases, with a total of 187 and 147 peaks in HOC and LOC, respectively, when using an R^2 of 0.4 (Figure 4.10A, top) for the maSigPro models, and 119 and 89 when using an R^2 of 0.6 (Figure 4.10B, top). Both R^2 profiles mimic the HOC and LOC phases in a similar way.

When grouping into three clusters we recapitulated the traditional functional modeling of the YMC, identifying OX, RB and RC phases. When using an R^2 of 0.4, a total of 72, 139 and 123 peaks were grouped in OX, RB and RC, respectively (Figure 4.10A, bottom), and the distribution changed to 110, 39 and 59 when the set of peaks obtained with an R^2 of 0.6 or higher (Figure 4.10B, bottom) was used. The most significant difference observed in comparing the profiles of peaks using an R^2 of 0.4 or 0.6 was the change in the RB phase, which shifted from having an early RB peak with $R^2 >=$ 0.4 to a late RB peak with $R^2 >=$ 0.6. The number of genes included in the RB phase greatly varied, suggesting that it might be the phase with the highest dispersion. When using an R^2 of 0.4, the RB peak is highly similar to the OX peak, whereas when an R^2 of 0.6 is used, the peak shifts towards the RC peak, meaning that some of the profiles could be redistributed to other phases depending on the R^2 used for the analysis. Using an R^2 of 0.4 allows for the detection of a higher number of oscillating profiles compared to using an R^2 of 0.6, and, although these profiles distribute in three clusters, the peaks corresponding to RB display a similar profile that of OX peaks, with a shifted peak and longer amplitude.

4.3.4 Detection of Transcription Factor binding sites

4.3.4.1 Binding sites detection with Wellington

Peak quantification and temporal modeling identified which peaks presented an oscillatory pattern across the YMC, allowing for capture of the changes in chromatin structure. However, the analyses fell short in determining the functional capabilities of these chromatin changes, such as gene expression regulation



A ATAC clusters R² = 0.4

HOC: 187 LOC: 147 OX: 72 RB: 139 RC: 123

B ATAC clusters R² = 0.6

HOC: 119 LOC: 89 OX: 110 RB: 39 RC: 59

Figure 4.10: ATAC-seq differential expression and clustering results. The DE ATAC peaks were obtained using R^2 thresholds in maSigPro of 0.4 (A) and 0.6 (B) and the resulting peaks were clustered in two (top A and B) and three clusters (bottom A and B) to compare the distribution of the peaks across the cycle.

due to changes in Transcription Factor (TF) binding. We thus proceeded to analyze the detected peaks and determined which TFs could be binding to chromatin.

To this end, peaks were subjected to footprinting using wellington [145], which detected the binding sites of TFs within the peak. Table 4.2 summarizes the total number of binding sites detected in each sample, including the 5 samples discarded from the dataset after quality assessment. Since samples from this dataset displayed generally low coverage, we analyzed whether grouping samples would help in the detection of TF binding sites.

Table 4.2: Total number of detected binding sites at each sample using Wellington.

Sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Wellington Sites	3493	2638	4119	2754	2378	1610	1188	1579	831
Sample	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Wellington Sites	2907	304	56	1345	1575	3568	3169	2932	2641

4.3.4.2 Grouping samples to define consensus Wellington regions

Peaks were subjected to binding analysis using each sample separately and combining samples in groups of two and three. Only consecutive samples were combined, thus representing similar chromatin states (e.g., Sample 2 is combined pairwise with samples 1 and 3 separately, and in three sample groups with samples 1 and 3, and 3 and 4). Figure 4.11A shows the number of binding sites when using samples independently, pairing them or grouping them in threes. Although the mean peak number did not vary much, dispersion in independent samples was much higher than when they were paired or combined in groups of three, most likely due to the impact of samples 10 and 11, which showed a low number of peaks.

Figure 4.11B represents the number of called binding sites per sample; the boxplot displays the number of sites from one sample alone and in combination with the consecutive samples in groups of two and three total consecutive samples. This boxplot evidences that samples 10 and 11 affect greatly the total

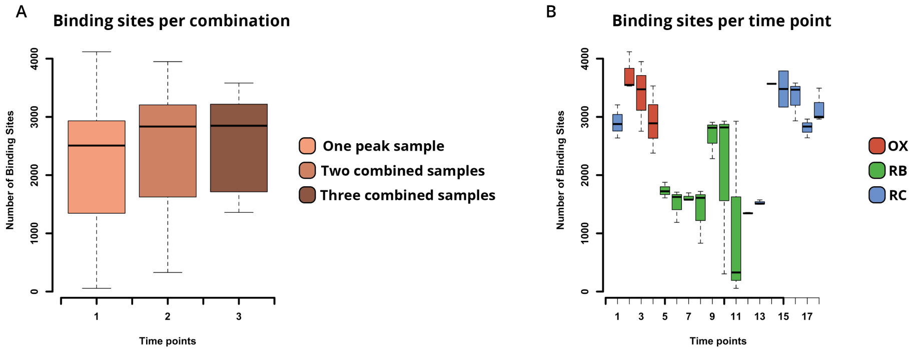


Figure 4.11: (A) Binding sites detected with Wellington when using mapping files from one sample or from two or three samples combined. (B) Distribution of number of binding sites per timepoint when using one, two or three combined mapping files that include that sample in Wellington for binding site discovery.

number of binding sites. It seems that OX and RC phases have a higher number of detected binding sites than RB. Since grouping samples did not have a profound effect on the number of detected binding sites, compared to analyzing samples separately (except for samples with low number of called peaks, which are excluded from the analysis), we decided to continue from this point with binding sites called from the individual analyses of the samples.

4.3.5 Functional description of TFs

4.3.5.1 TF detection with motifs

After detection of binding sites in each peak, binding sites were subjected to TF motif finding. Nucleotide sequence from all binding sites were extracted and scanned against JASPAR database using FIMO software [62], which extracted motifs from the DNA input sequences and revealed the TFs that bind to each motif.

TFs were associated with target genes by proximity, and we used RG-match software [56] to identify gene targets and filter only for genes that bind to promoter regions. Table 4.3 shows the TFs with the overall highest number of binding points. Next, we asked whether these TFs bind differentially across the cycle, and if they regulate a different set of genes, depending on the YMC phase.

Table 4.3: Top 8 Transcription Factors (TFs) with the overall highest number of binding sites. This table displays the number of binding sites for each TF in all time points.

TF	T1	T2	T3	T4	T5	T6	T7	T9	T14	T15	T16	T17	T18
AZF1	137	73	180	69	83	78	24	54	104	124	114	89	85
SUM1	89	65	151	39	37	49	18	37	74	116	81	63	60
ABF1	91	53	110	59	46	50	35	36	67	97	66	89	78
EDS1	72	47	82	40	40	46	22	25	55	57	46	72	50
SUT1	60	52	58	41	42	48	14	36	56	57	64	43	53
YPR196W	53	38	63	21	45	37	20	31	53	57	48	60	48
MIG2/3	57	47	52	46	47	32	19	30	34	52	44	43	53

We used Tucker3 to decompose the 3-way binding matrix in three modes, namely genes, TFs and time points [28]. The matrix was constructed as a binary data matrix that combines TFs (X mode) with their target genes (Y mode) in a time-dependent manner (Z mode). We observed that although most of the genes do not separate (Figure 4.12A), some TFs involved in similar processes did group together, such as Mig2, Mig1 and Sut1, which are involved in sterol and fatty acid metabolism, and Rsc3 and Abf1, which are chromatin remodelers. Azf1, which has glucose sensing activity, separated from the other TFs (Figure 4.12B). Inspecting model projection for Mode 3 we concluded that the time-course information was partially captured; specially, we observed separation of OX and RC from RB in the first component (Figure 4.12C).

4.3.5.2 Functional description of TFs

Although multiple TFs were detected through this analysis, Azf1 is particularly interesting, not only due to the high number of regulated genes, but also because it is an important regulator of genes that respond to glucose to mediate diauxic shift. SUM1 is involved in mitotic repression, and Abf1 and Rsc3 mediate chromatin remodeling. Mig2, Mig3 and Sut1 are involved in fatty acid metabolism, while Eds1 and Rgt1 are involved in glucose metabolism.

In the next chapter, we will combine these TFs with their regulated genes and perform an integrated analysis to gather information regarding their impact in the different parts of the Yeast Metabolic Cycle.

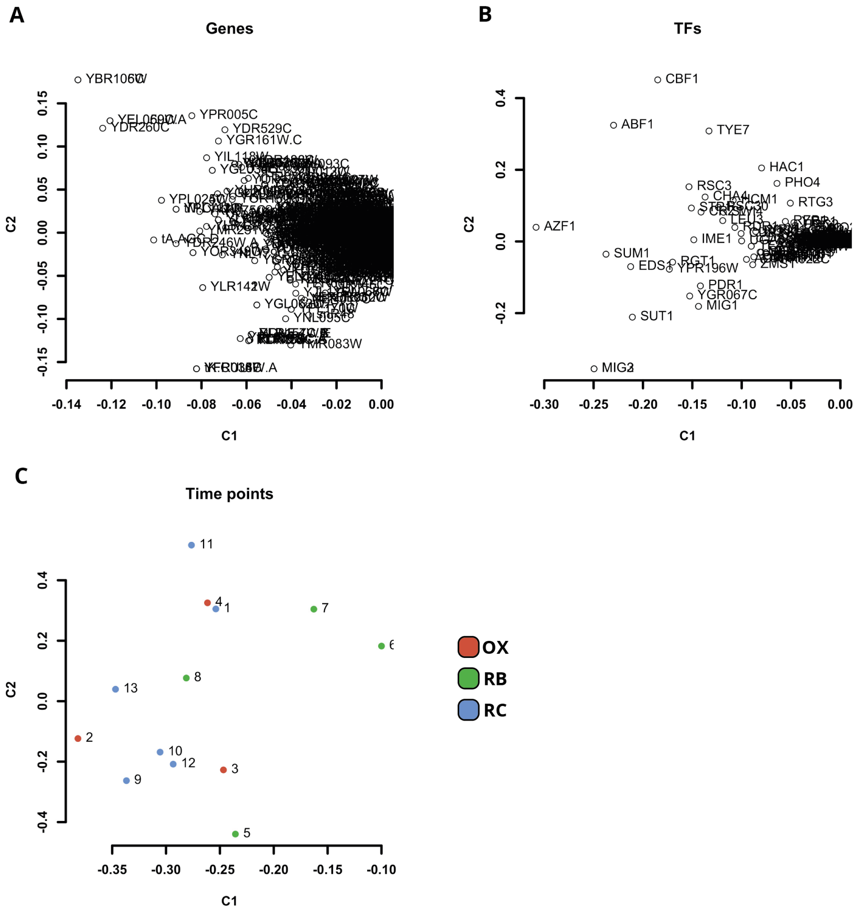


Figure 4.12: Tucker3 in the three-dimensional transcription factor binding matrix allows for separation of the three modes in genes (A), transcription factors (B) and time points (C).

4.3.6 NET-seq data analysis

NET-seq data obtained from [51] was analyzed by maSigPro, using a similar approach to that used for the metabolomics and ATAC-seq data but setting a fixed minimum R^2 of 0.6 or higher. The resulting genes were subjected to a PCA analysis (Figure 4.13A) that showed that the flow of the YMC was captured by the two first PCA components. The two components explained 66.7% and 20.7% of the data variability, respectively.

K-means clustering of differential genes recapitulated the three YMC phases

(Figure 4.13B). OX, RB and RC phases contained 373, 232, and 564 genes, respectively, indicating that the NET-seq data clearly defined the three phases. As in ATAC-seq, RB phase showed a lower amplitude than OX and RC phases, while these last two phases had completely opposite patterns.

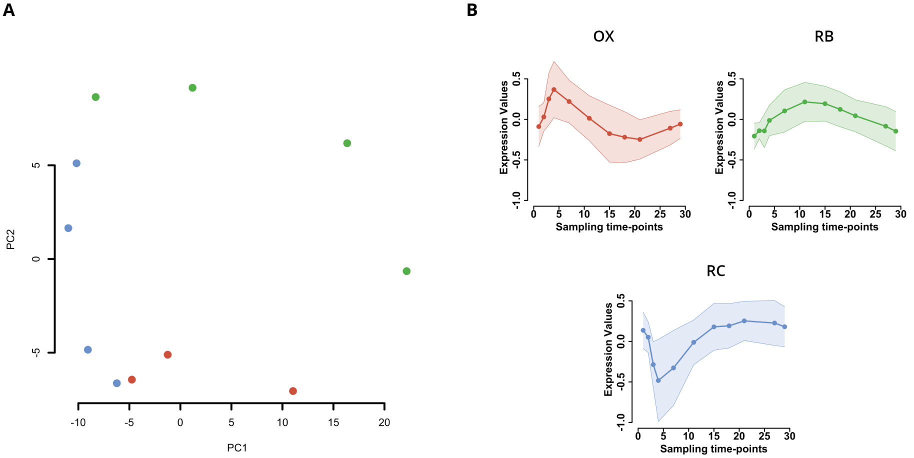


Figure 4.13: PCA of NET-seq differentially expressed genes (A) and result of clustering in three clusters after differential expression (B).

4.4 Discussion

The Yeast Metabolic Cycle (YMC) is characterized by an orchestrated regulation of gene expression by metabolic and epigenetic oscillations. In order to address the mechanisms by which these three molecular layers cooperate to define the periodic rhythmicity of the cycle, the combination of multi-omics information with multi-layered statistical modelling is required. Although histone modifications and gene expression have been well studied in the YMC context, the metabolomics dataset available are scarce and cover the cycle with low resolution. Moreover, the measurement of chromatin architecture in this system has not been previously done.

In this chapter we obtained comprehensive metabolomics and an ATAC-seq dataset of the YMC and processed a recently released NET-seq dataset [51]. The three datasets complement the RNA-seq and histone modification datasets analyzed in the previous chapter. The metabolomics dataset was obtained in quintuplicates, with 21 matching time points in each replicate. A total of 60 metabolites were obtained from five different chemical extractions that included organic acids, amino acids, reduced nucleotides, oxidized nucleotides and Acyl-CoAs. The 21 matching time points were aligned, and the PCA of the resulting matrix revealed that the samples recapitulated the three YMC phases and followed a temporal order of extraction.

Differentially abundant metabolites were divided in two or three clusters depending on the R^2 threshold specified in the maSigPro analysis. The restrictive division of the metabolomics profiles in two clusters revealed that the profiles followed two clusters, defined as High Oxygen Consumption (HOC) and Low Oxygen Consumption (LOC) phases; these profiles separate the YMC into two phases according to the levels of dissolved oxygen. If less restrictive differential expression parameters were used, the metabolic profiles could still be separated in two clusters, but the separation in three clusters seems to better capture the oscillatory changes of metabolite abundance during the cycle. These three clusters match those described for gene expression in previous studies, but with a slight delay in the apparition of each phase's peak. The types of metabolites included in each cluster are in agreement with the function of the genes of the same cluster. For instance, NTPs or Malonyl-CoA present in the RC metabolic cluster correspond to a high energy state that matches the fatty acid degradation seen in RC phase gene expression. This high energy state is expected at the end of RC phase, when fatty acid degradation has concluded, and the delay in the profile's peaks compared to RNA-seq profiles points to a late accumulation of metabolites in all phases. The functional interpretation of metabolic activity in the mark of gene expression functionalities is limited at this stage, since both datasets were analyzed independently. In the next chapter, we address their integrative analysis.

One question we addressed here is whether the dynamics of metabolic changes are best represented by a 2 or by a 3 phase model, i.e., if clustering into 2 or 3 groups was better. The answer to this question is not straightforward, as both R^2 maSigPro analyses seem to fit the two clustering approaches well. However, the separation of the metabolites into three clusters did accurately model the oscillations that occur during the cycle, with OX phase as the most variable cluster that gets redistributed when the clustering is set to two clusters.

As for the ATAC-seq dataset, after quality assessment of the sampling, five samples were discarded from the study due to their low read and mapping quality. Note that these five samples corresponded mainly to the late RB phase; this could point to the difficulty of subjecting RB phase cells to chromatin digestion, maybe due to problems isolating nuclei or a thicker cell wall or because the start of the cell cycle (reported at this stage of the YMC [20]) hinders the activity of transposase. After these samples were discarded, PCA exploratory plots showed that samples group according to their phase, suggesting that the samples that were not discarded captured the chromatin changes that happen during the YMC. ATAC-seq peak profiles could be clustered in three phases, although RB phase is the least defined since it has a similar pattern to OX phase. Although the arrangement of chromatin accessibility in two phases matches with the results shown in the integration of RNA-seq and histone modifications in the previous chapter, we do not exclude that this pattern could be caused by the low sampling resolution of RB phase after corrupted samples were discarded.

ATAC-seq data was also analyzed to identify the accessibility of TFs to chromatin. We obtained a long list of TFs that bind to the chromatin-occupied regions across the whole cycle, with different binding depending on the sampling time point. Azf1 and Sum1 stand out as the TFs with highest number of regulated genes, although the presence of Abf1 and Rsc3 chromatin remodelers is especially interesting. Ino80, an ATP-dependent chromatin remodeler, has been recently associated with the YMC cycle progression [61], and it would be interesting to study if other chromatin remodelers are associated to Ino80 activity or involved in shaping other YMC phases. All in all, most TFs were involved in metabolic activity, and their role in regulating gene expression will be studied in the next section to further describe the processes they regulate.

NET-seq profiles followed the established three YMC phase distribution, where RB phase had an arguably low amplitude and lower number of genes, which may suggest a lower impact than OX and RC phases. On the other hand, OX and RC presented antagonistic profiles that reflect the two most transcriptionally active intervals during the cycle.

Although here we analyzed the dynamics of metabolite changes, the NET-seq phases and the TFs that bind to a large number of chromatin regions, the best way to answer the question of which metabolites have a higher impact in the cycle or which TFs have a higher influence in gene expression is to integrate the information from the three datasets with gene expression changes. This question is studied in the next chapter.

Chapter 5

Multi-omics integration of the Yeast Metabolic Cycle

5.1 Introduction

Biological rhythms are key features of eukaryotic cells characterized by transcriptional cycles at specific time lapses. These coordinated oscillations in the expression of thousands of genes have been studied in multiple systems as ultradian or circadian rhythms [97, 187], and their regulation has been linked to changes in the cell epigenetic landscape and in its metabolic fluxes [21, 126]. The Yeast Metabolic Cycle (YMC) is a model system for the study of biological rhythms. *S. cerevisiae* follows continuous cycles of metabolic activity when cultured under aerobic, nutrient-limited conditions. While the YMC is characterized by two distinct phases of oxygen consumption: High Oxygen Consumption (HOC) and Low Oxygen Consumption (LOC) [126], genes typically change their expression according to three functional phases: Oxidative phase (OX), Reductive Building phase (RB), and Reductive Charging phase (RC) [187]. These three functional phases are also recapitulated when looking at changes in histone modifications and metabolites [102, 188].

The control of gene expression has been traditionally understood as the result of Transcription Factor (TF) activity [103, 175]. However, recent research has used the YMC as a model to study the impact of chromatin regulation on gene expression [61] and the coordination between epigenetic changes and transcriptional activity during biological rhythms, suggesting a “metabolism-epigenome-transcriptome loop in the YMC” [102]. Oscillations in metabolite levels during the YMC suggest that accumulation of key compounds could contribute to the coordination of the cycle [188], but no study has analyzed together the metabolism-epigenome-transcriptome cross-regulation of the YMC.

Biological rhythms have been extensively studied using high-throughput molecular assays. Previous research has either focused on a single molecular layer [5, 76, 99] or has combined two omic modalities, for example, gene expression and histone modifications [1, 23, 90], gene expression and metabolism [42, 158, 207] or gene expression and TFs [40, 176]. Studies that integrate a higher number of regulatory layers are lacking. This is probably due to the difficulty of gathering multi-omic information from the same biological system and to the complexity of a multi-layer statistical analysis. One of the challenges when modelling the influence of metabolic changes in biological rhythms is finding the appropriate analysis methodology that can reveal the sequential steps and/or feedback loops that operate in the control of these processes. In this chapter, we use the YMC to deploy a novel analytical strategy for multi-omics data integration that addresses the particularity of the multi-layer modelling of biological rhythms. This strategy revealed which metabolites are likely to controlling the epigenetic oscillations that eventually contribute to the regulation of gene expression.

The multi-omics data integration strategy presented here uses multivariate methodologies based on latent variables to study the inter-relationship between different molecular layers. We used the multi-omics dataset described in previous chapters, which includes ChIP-seq of several histone marks, RNA-seq, NET-seq, ATAC-seq and metabolomics. We applied Partial Least Squares – Path Modelling (PLS-PM) to model the flux of molecular information that connects metabolites with gene expression through epigenetic changes in the YMC. To our knowledge, this is the first time that PLS-PM has been used to model biological rhythms. We found that regulation of gene expression mostly take place at OX and RC phases, while the RB phase is not affected much by epigenetics oscillations. Active TFs detected with ATAC-seq are also expressed in a two-phase mode, which suggests that RB phase is subjected to different regulatory mechanisms other than OX and RC. Finally, we identified a metabolite signature that is highly correlated with the dynamics of epigenetic changes, as well as group of compounds that may accumulate as enzymatic products of proteins coded by periodically expressed genes.

5.2 Methods

5.2.1 Datasets

We complied a multi-omics dataset by joining the data from RNA-seq and ChIP-seq of two histone modifications (H3K9ac and H3K18ac) from [102] presented in the first chapter of this thesis, the NET-seq dataset obtained from [51], and the ATAC-seq and metabolomics data described and pre-processed in the second chapter of this thesis. Data preprocessing was done as described in previous chapters. Additionally, RNA-seq data were subjected to TMM normalization for the analyses described here. All the datasets span the totality of the Yeast Metabolic Cycle and, when combined, form a multi-layer dataset that covers the metabolome-epigenome-transcriptome axis of the cycle.

5.2.2 Differential Expression and Clustering

Each omic feature was modeled with a polynomial function of time by applying the linear regression method implemented in the maSigPro R package [138]. We set a maximum polynomial degree of 3 to allow for linear, quadratic and oscillating patterns. Differentially expressed (DE) omic features were those with a significant model (FDR adjusted p-value < 0.05) and a minimum R^2 value. This minimum R^2 value was 0.6 for gene expression, histone modifications and NET-seq, and 0.4 for ATAC-seq and metabolomics to accommodate a higher dispersion and different sampling intensities of these datasets (see chapter 4 for the evaluation of R^2 thresholds in these omics modalities). For each omic, DE features were scaled and clustered with the k-means algorithm into three clusters to match the number of phases in the YMC. Specifics of DE and clustering are further described in chapters 3 and 4.

5.2.3 Matching omics datasets

The different omics datasets combined in this study were extracted from different YMC experimental runs and had slightly different positions in the cycle (Figure 5.1). For integrative analyses, alignment of YMC data-points across omics datasets was required. We created a function that related oxygen levels to YMC time points and used it to decide which sampling points should be averaged, removed or interpolated at each omics modality in order to have a uniform set of data points across all omics types.

Combining all datasets reduced the number of aligned time points across all datasets to 11, risking loss of predictive power. To address this limitation, each -omic was modeled as a function of time using spline regression models in R and these models were used to predict the missing time points of each data type, resulting in a dataset that contained a total of 30 coincident time point for each omic modality. This complete 30 time-points dataset contains enough temporal data for an omics integration analysis using PLS-PM (section 5.2.6).

5.2.4 Partial Least Squares Regression

Partial Least Squares Regression (PLS) is a multivariate regression method based on dimension reduction. PLS identifies a set of latent variables, also called components, that summarizes the information in both predictor ($\mathbf{X}$) and response ($\mathbf{Y}$) matrices in such a way that the covariance between $\mathbf{X}$ and $\mathbf{Y}$ components is maximized. Opposite to classical regression models, PLS benefits from multicollinearity in the predictive dataset and can work with more predictors than observations.

In this study, PLS was applied to model gene expression as a function of metabolomics, transcription factor activity and the two histone modification datasets. Let $\mathbf{Ynxp}$ be the gene expression data matrix with n samples and p genes, and $\mathbf{X}_{nxm}$ the matrix representing either metabolites abundance, ATAC-seq peaks intensities or ChIP-seq quantification with n samples and m features. The PLS model is established as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{5.1}$$

Where $\mathbf{B}$ is the matrix of regression coefficients and $\mathbf{E}$ the residuals matrix. The underlying component-model of PLS is:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{H} \tag{5.2}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{5.3}$$

Where $\mathbf{P}$ and $\mathbf{Q}$ are the loading matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively; $\mathbf{T}$ and $\mathbf{U}$ are the score matrices; and $\mathbf{H}$ and $\mathbf{F}$, the residuals. $\mathbf{T}$ ($\mathbf{U}$) and $\mathbf{P}$ ($\mathbf{Q}$) provide the projections of the original observations and variables (time points and omic features in our case) onto each of the components of $\mathbf{X}$ ($\mathbf{Y}$) in the new, low dimensional space.

$\mathbf{X}$ and $\mathbf{Y}$ matrices were centered and scaled for this analysis and a PLS model with two components was applied, as this was sufficient to summarize the behavior of the YMC. We used R^2 and Q^2 calculations to estimate the explained and predicted variance of the model, respectively.

5.2.5 Generalized Linear Models

The MORE package (https://github.com/ConesaLab/MORE) was used to fit Generalized Linear Models (GLMs) that predict, for each gene, its expression (response variable *y*) as a function of the effects of a number *R* of regulatory features (explanatory variables x_1,. . . , x_R) given by the multi-omics measurements. Since we assume gene expression follows a Gaussian probability distribution, the regression model for each gene to fit is:

$$y = \beta_0 + \sum_{i=1}^{R} \beta_i x_i + \epsilon \tag{5.4}$$

Where y are the gene expression values, β_i the regression coefficients,and ϵ the error term. TF expression values were identified by ATAC-seq footprinting analysis, wherein the TF motif was found at the promoter region of each gene, and histone modification values are the ChIP-seq peak quantification values for histone modifications at the promoter region of each gene. Significant regulators were those with significant associated regression coefficients (p-value $<$ 0.05).

5.2.6 PLS Path Modeling

PLS Path Modeling (PLS-PM) is a component-based method for estimating Structural Equation Models (SEMs). SEM is a methodology for estimating and testing a network of relationships between variables that can either be measured or unobserved. The unobserved (or latent) variables represent abstract concepts of the data domain that cannot be measured directly but can be indirectly inferred from the measured data. PLS-PM estimates these latent variables (LVs) as a function of the measured variables and infers relationships between LVs by fitting linear regression models. Regression models are initially defined based on theoretical relationships among abstract concepts (LVs) given previous knowledge about the biological system under study. PLS-PM identifies the significant regression coefficients, and hence, the relationships that are sustained by the data. These relationships among LVs can be represented in the form an acyclic network called a “path model”.

The path model has two sub-models: a measurement model and a structural model. The measurement or outer model connects the observed data (indicators or manifest variables) with their associated LV, which is calculated as a weighted sum of its indicators (see Equation 5.5 for LV Y_j).

$$Y_j = \sum_k w_{jk} X_{jk} \tag{5.5}$$

Where X_{jk} is the manifest variable k associated to LV *j*, and coefficients w_{jk} are called weights, which are estimated with an iterative procedure [160]. The X matrix containing the observed data must have been previously scaled.

The structural or inner model connects the LVs by multiple linear regression (see Equation 5.6 for LV Y_j).

$$Y_j = \beta_0 + \sum_i \beta_i Y_i + \epsilon_j \tag{5.6}$$

Where the subscript i refers to all the LVs that potentially predict Y_j, β_i are the path coefficients and ϵ_j is the error term.

The PLS-PM algorithm calculates the inner and outer models through an iterative procedure where LVs and their indicators are estimated together with the coefficients of the regression models between LVs. This process starts with the model definition (setting possible connections between LVs) and follows rounds of outer model and inner model validations.

5.2.6.1 Model definition

We used the plspm package in R [161] to define the elements of the outer (LVs or path nodes) and inner (path connections) models. Path nodes (LVs) were initially defined for each omics modality and phase of the YMC. The indicators, or observed variables, associated to each LV was the set of maSigPro significant variables belonging to each of the clusters and assigned to a YMC phase (Figure 5.2). Therefore, the initial model was set with a total of 19 (unobserved) LVs that corresponded to gene expression (measured by RNA-seq), TFs (measured by RNA-seq and identified by ATAC-seq), transcription (measured by Net-seq), histone modifications (measured by ChIP-seq) and metabolites (measured by metabolomics) at each YMC phase. Formally, the model used matrices $\mathbf{X}^{a,b}_{a_j,t}$, where the matrix $\mathbf{X}^{a,b}_{a_j,t}$ contained the a_j significant features of omics a $\in$ $\{$RNA-seq, RNA-seq (TF), ChIP-seq, Net-seq, Metabolomics$\}$ at phase $b \in$ $\{$Ox,Rb,Rc$\}$, and t = 30 is the total number of data-points in the YMC. The path model was initially set according to theoretical knowledge of their molecular regulatory connections, where metabolite levels influence histone modifications and histone modifications influence gene transcription.

Once outer and inner models were defined, the LVs were calculated from their corresponding set of indicators in a reflective way, that is, indicators were considered to be caused by LVs (see Equation 5.7 for LV Y_j).

$$X_{jk} = \lambda_{0jk} + \lambda_{jk} Y_j + \epsilon_{jk} \tag{5.7}$$

Where λ_{0ji} is an intercept term, λ_{ji} are the loadings, and ϵ_j is the error term for each LV Y_j ($j = 1, ..., 19$) and its corresponding indicators X_{jk} ($k = 1, ..., a_j$)

5.2.6.2 Outer model validation

Reflective indicators must satisfy some validation criteria to meet the requirements of the outer model, such as unidimensionality, meaning that all indicators are supposed to be positively correlated. Additionally, indicators should be sufficiently explained by their latent variable. This means that, during the PLS-PM calculation procedure, indicators that do not meet these requirements are excluded from the final model. This is an iterative process: every time the model is re-calculated, indicators are evaluated until all remaining indicators satisfy the validation requirements. Indicators in a LV with a loading value lower than 0.7 were discarded. Additionally, the unidimensionality criteria needs to be maintained. Unidimensionality of a LV indicates whether the indicators that belong to a LV are highly correlated and have the same trend. To assess unidimensionality we used the Cronbach's alpha (C-alpha or α) [30] (Eq. 5.8) and the Dillon-Goldstein's rho (DG-rho or ρ) [161] (Eq. 5.9) implementations present in the plspm R package, with modifications to allow for data structures with more variables than observations.

$$\alpha_j = \frac{N\overline{c}_j}{\overline{v}_j + (N-1)\overline{c}_j} \tag{5.8}$$

Where N is equal to the number of indicators, $\overline{c}$ is the average covariance among the indicators and $\overline{v}$ equals the average variance of all the indicators.

$$\rho_j = \frac{(\sum_{k=1}^{a_j} \lambda_{jk})^2}{(\sum_{k=1}^{a_j} \lambda_{jk})^2 + (p_j - (\sum_{k=1}^{a_j} \lambda_{jk})^2)} \tag{5.9}$$

Where λ_{jk} is the factor loading of the indicator k and LV_j, and p_j is the number of indicators in LV_j.

We imposed a threshold of 0.7 in both measurements to determine that the LV was explaining its block of indicators well.

Finally, cross-loadings were calculated to assess the accuracy in the indicator assignment to the correct LV. We expect the loading of all indicators to be higher than their cross-loadings, since the opposite would indicate that the indicator has a higher correlation with the value of a different LV than its own. We calculated the cross-loadings of all indicators and compared them to their loadings to evaluate whether the LV assignment was correct.

5.2.6.3 Inner model validation

Once the outer model was validated and redefined to meet the required assumptions, the inner model was estimated, that is, the values of the path coefficients connecting the LVs were obtained. The quality of inner model was assessed using different criteria.

To evaluate the quality of the connections between LVs, we obtained the coefficients of determination (R^2) of the regression models and the average redundancy. The R^2 coefficient indicates the percentage of variance of a LV explained by its connected LVs. The redundancy represents the percentage of variance of an indicator in a LV explained by the independent LVs in the corresponding regression model (equation 5.10).

$$Rd(LV_j) = \overline{\lambda}_j R_j^2 \tag{5.10}$$

Where $\overline{\lambda}_j$ is the average of the loading for indicators in LV j and R_j^2 are the coefficients of determination for LV_j.

The statistical significance of the inner regression model parameters was further estimated by a bootstrapping procedure, where M samples are created in order to obtain M estimates for each parameter in the PLS model. Each sample is obtained by sampling with replacement from the original dataset, with sample sizes equal to the number of cases in the original dataset. Confidence intervals from the bootstrapping can be used to validate the significance of the relationships using a non-parametric approach.

The significance of the inner model was also evaluated by comparing direct and indirect effects. Indirect effects refer to the influence of one LV on another LV to which is not directly connected and are obtained as the product of the path coefficients by taking an indirect path. An indirect effect higher than the direct effect could suggest a missed direct connection between the two LVs.

5.2.6.4 Global model evaluation

The overall performance of the model was measured using the Goodness of Fit (GoF), which determines the model prediction power. GoF is a pseudo-Goodness of Fit measure that attempts to account for the overall quality of both the measurement and the structural models. GoF assesses the overall prediction performance of the model by taking into account the communality and the

R^2 coefficients, so it is a compromise between the quality of the measurement model and the quality of the structural model:

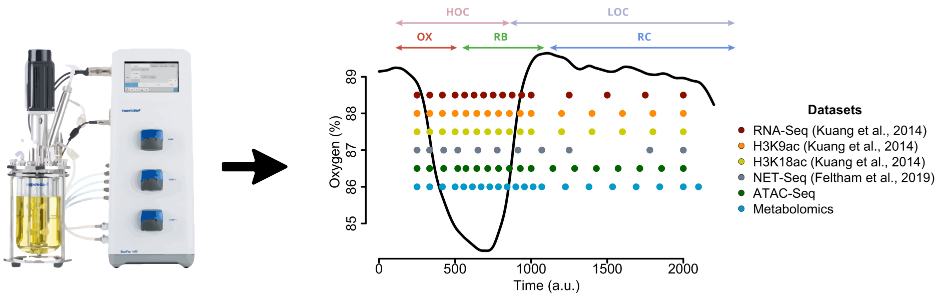
$$GoF^2 = \overline{Comm} \times \overline{R^2} \tag{5.11}$$

Where $Comm$ refers to the communality and R^2 to the coefficients of determination. A GoF higher than 0.7 is generally considered good. Finally, we evaluated the quality of the proposed inner model using a permutation approach. Basically, using the set of LVs in our model, we randomly permuted the relationships among them. The process was repeated 1000 times and the Goodness-of-Fit (GoF) of our model was compared to the GOF of the random models. When building random models, we imposed that LVs had a minimum of one relationship and a maximum of 3. Models where all relationships for a given LV were deleted had to be discarded, this process was repeated until we had a total of 1000 valid models.

5.3 Results

5.3.1 Multi-layered omics dataset

Statistical integration of high-throughput data in biological rhythms requires datasets that cover the totality of a cycle with high resolution [144]. In this study we have used five different datasets that cover the metabolic, epigenetic and gene expression dynamics of the YMC. Each dataset was obtained from a different YMC cycle, and sampling times varied from one dataset to another. We made use of oxygen profiles to align the samples across omic datasets. Figure 5.1 presents the approximate sampling points of the datasets used, while Table 5.1 indicates how samples were interpolated, averaged or discarded to reach uniform alignment across datasets.



Controlled and stable conditions

Even sampling accross the different phases of the YMC for multiple omics datasets

Figure 5.1: Sampling scheme for the multi-omics YMC dataset

The application of PLS Path Modeling requires a sufficient number of observations. To avoid losing sampling points when matching data matrices, spline regression models were used to predict missing points in each of the omics measured. The application of spline models allowed for the acquisition of datasets for each omic that contained 30 time points, allowing the application of the multivariate strategies presented in this work to study omics cross-regulation. Table 5.2 provides details on the time points that were inferred using spline linear models for a better alignment of the data in the PLS-PM model.

Table 5.1: Time point modifications made in order to align the datasets used in the omics integration analyses.

Omic 1	Omic 2	Modifications omic 1	Modifications omic 2
RNA	H3K9ac and H3K18ac	Averaged time points 9 and 10	Averaged time points 12 and 13
RNA	Metabolites	Averaged time points 9 and 10	Averaged time points: 2, 3 and 4; 11 and 12; 13 and 14
RNA	ATAC-seq	Averaged time points: 2 and 3; 5 and 6; 7 and 8; 9 and 10; 12 and 13	Averaged time points: 2 and 3; 11 and 12; 13 and 14
NET-seq	Metabolites	No changes	Averaged time points: 7, 8 and 9; 10, 11 and 12; 13 and 14; 15 and 16; 19 and 20; 21 and 1. Deleted time points: 2 and 3

Table 5.2: Time points inferred in order to produce a complete dataset combining all omics in a multi-omics analysis.

Omic	Inferred Timepoints
RNA-seq	7, 9, 11, 13, 15, 18, 19, 20, 22, 24, 25, 27, 28 and 30
ChIP-seq	6, 8, 10, 12,14, 15, 18, 19, 21, 23, 25, 26, 28 and 30
Metabolomics	7, 10, 12, 14, 19, 21, 24, 26 and 27
NET-seq	5, 6, 8, 9, 10, 12, 13, 14, 16, 17, 19, 20, 22, 23, 24, 25, 26, 28, 30

5.3.2 Omics profiles suggest that both gene expression and metabolites oscillate in three phases.

Previous studies on the YMC have characterized gene expression changes in three well differentiated phases. We reproduced this analysis for all the omic modalities in this study by modeling each feature YMC value as a function of time using the polynomial regression implemented in the maSigPro package [26] and clustering features with similar temporal behavior with the k-means algorithm. We obtained three distinctive clusters in each case, and each cluster could be associated to a YMC phase (Figure 5.2). In the case of RNA-seq, genes within each cluster mostly coincided with genes previously reported for each YMC phase (Figure 5.3). Modeling and clustering of histone modifications into three groups also recapitulated the three YMC phases. Histone modifications in OX and RC phases followed the same profiles as gene expression data; howeber, a late peak in H3K9ac and a low number of genes (88 genes) for H3K18ac was observed in RB phase. The OX peak for both histone modifications, as for RNA, was the sharpest across phases (Figure 5.2).

Although NET-seq and ATAC-seq datasets could also be modelled in three phases, the oscillations at the RB phase in NET-seq had low dynamic range, while RB phase in ATAC-seq was shifted towards OX phase, suggesting the presence of two sub-clusters within the same profile. The shape of OX and RC phases from ATAC-seq and NET-seq data resembled their corresponding RNA-seq profiles. Finally, the three oscillatory phases that resulted from metabolite clustering were as well defined as they were for gene expression. Figure 5.2 shows that peak position of YMC phases obtained from metabolomics data is shifted to later timepoints compared to other omics. Moreover, the OX phase of the metabolomics dataset has a larger amplitude in comparison to the RNA-seq OX phase.

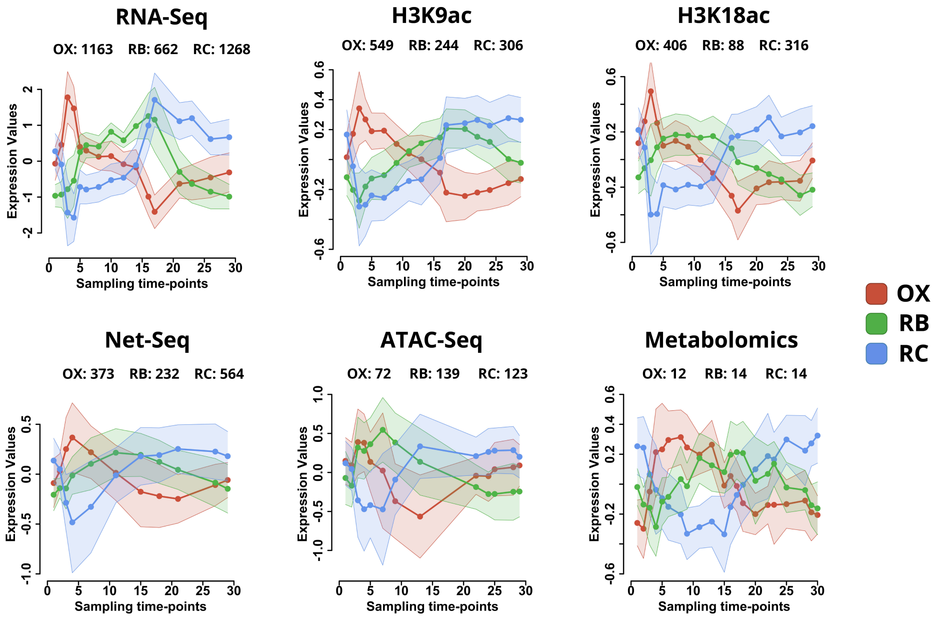


Figure 5.2: Profiles of YMC clusters across the studied -omics. RNA-seq presents three well defined clusters, while for H3K9ac and ATAC-seq, the RB phase is shifted to later and earlier time points, respectively. ATAC-seq and H3K18ac RB phase shows a low dynamic range and has a low number of members for H3K18ac. Metabolites present three phases that appear to be shifted to later time points compared to RNA-seq

All together, our analysis of omics data oscillations across the YMC suggests that all measured molecules can be modelled into the three phases, previously described for gene expression, which suggests a possible mechanistic coordination. However, although RB phase is well-shaped in the RNA-seq dataset, it exhibits the least consistency across other omics layers, presenting shifts, low amplitude or low representation, suggesting that transcriptional control in RB phase might not be as tightly subjected to chromatin control as in other phases of the YMC.

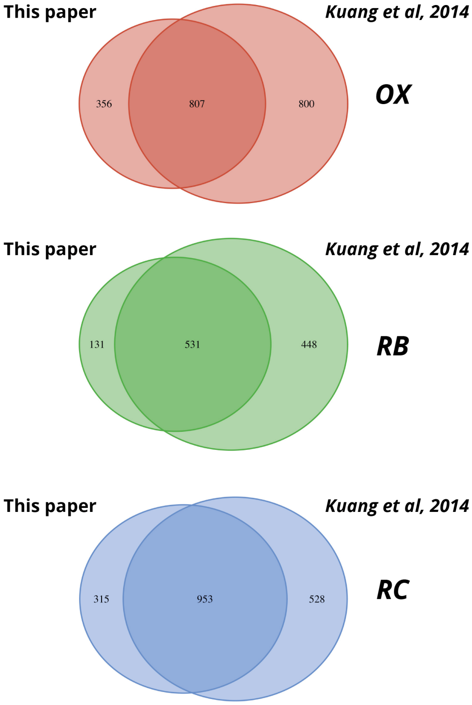


Figure 5.3: Comparison of genes in our clusters compared to the clusters from Kuang et al, [102].

5.3.3 Chromatin associates with gene expression at two phases while metabolism coordinates with three functional phases

We next asked how metabolism and chromatin dynamics coordinate with gene expression during the YMC, and whether all molecular layers have a similar impact on gene expression. We used Partial Least Squares Regression (PLS) [59] to answer this question, since this method is able to model the omics global response, capture interactions among omics datasets and find groups of features that may mediate these interactions. Since our goal is to understand the regulation of gene expression, we modelled RNA-seq data as a function of ChIP-seq data for histone modifications, metabolomics and ATAC peaks separately.

5.3.3.1 Histone modifications predict gene expression changes, especially at OX and RC phases.

PLS models of gene expression in response to chromatin marks presented a high explanation and prediction power, with R^2 values of 0.78 and 0.79 and Q^2 values of 0.64 and 0.67 for H3K9ac and H3K18ac, respectively. The loadings of the first two RNA-seq components showed a clear separation of the three YMC phases (Figure 5.4), which were also recapitulated by the ChIP-seq loadings. In agreement with results from the previous analyses, the histone mark data for genes labelled as OX and RC-associated, showed the greatest separation in the first component, while RB genes were not as discriminated. We further studied the significance of the connection between gene expression and histone modifications using the MORE tool [181], which adapts Generalized Linear Model (GLM) regression to the multi-omics scenario. In this case, gene expression was modelled as a function of both histone modifications, and regulatory relationships were concluded from the significance of the regression coefficients in the GLM. Genes with significant regulation by either H3K9ac or H3K18ac were marked in black in Figure 5.4. This analysis confirmed that H3K9ac and H3K18ac have a higher impact on OX and RC phases, respectively, while none of the marks showed a strong connection to gene expression at the RB phase. We concluded that H3K9ac and H3K18ac are transcriptional activation marks in OX and RC phases, but not in RB phase (Figure 5.4C).

5.3.3.2 Metabolic changes have a strong correlation with the three gene expression phases

Next, we studied the effect of metabolic oscillations on nascent (NET-seq) and mature transcripts (RNA-seq) by fitting PLS models wherein metabolomics data was used as the explanatory variable and gene expression was used as the response variable (Figures 5.5 and 5.6 for RNA-seq and NET-seq, respectively).

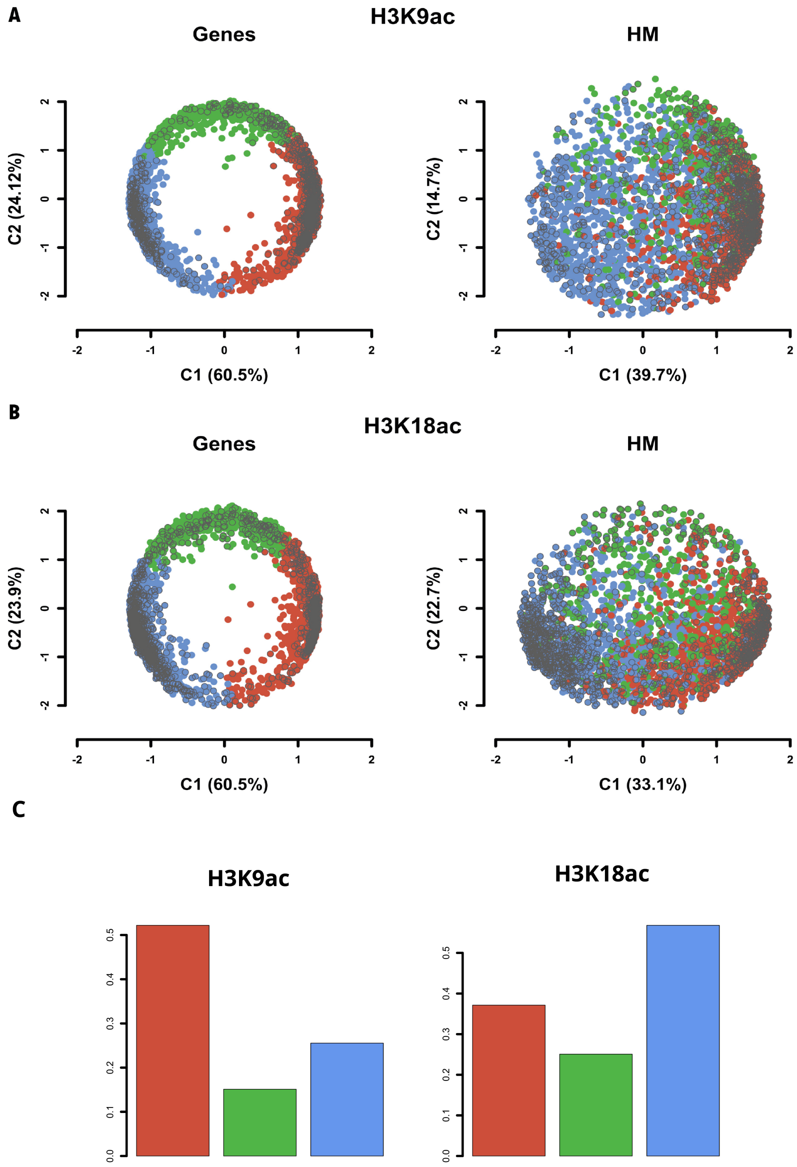


Figure 5.4: PLS of histone modifications (explanatory variable) and gene expression (response variable). Both histone modifications explain gene expression variability. (A) H3K9ac has a higher link with OX genes, and (B) H3K18ac is more related to RC genes. (C) Percentage of genes from each phase found significant for H3K9ac and H3K18ac MORE models.

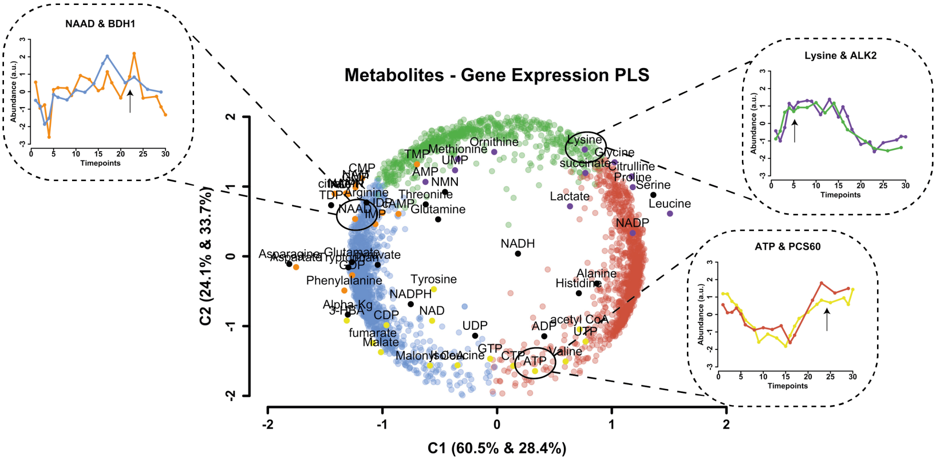


Figure 5.5: Metabolites influence on gene expression. PLS between metabolites (explanatory variables) and gene expression (response) indicates accumulation of metabolites throughout the following gene expression.

The first component accounted for 60.5% and 57% of the metabolomics data variation in RNA and NET PLS models, respectively. The PLS model with 2 components achieved a high prediction power for RNA-seq data with an R^2 value of 0.8 and a Q^2 value of 0.7. NET-seq PLS had an R^2 value of 0.76 and a Q^2 value of 0.67. The two-component PLS models recapitulated the phase separation and showed a time-sequential distribution of features, which is depicted in Figure 5.5. Metabolite projection into the lower dimensional space indicated accumulation of compounds at the transition between the gene expression phases, corroborating the clustering pattern observed in Figure 5.2.

Detailed analysis of metabolite projections revealed different groups of metabolites at each phase, suggesting early and late accumulation of metabolites in the YMC phases. The late OX-phase showed an accumulation of amino acids, late RB phase showed an accumulation of nicotinamide (NAM)-derivative metabolites and nucleoside monophosphates (NMPs), and early RC phase showed accumulation of aspartate, phenylalanine and GDP. The late RC phase was characterized by accumulation of TCA cycle intermediates and Malonyl-CoA, and early OX phase showed accumulation of nucleoside triphosphates (NTPs) and acetyl-CoA (Figure 5.5). We concluded that metabolites from each cluster accumulated in between RNA-seq phases in two groups, suggesting an early and a late accumulation of metabolites for each phase. Each metabolic cluster contained groups of metabolites that were associated to common biological processes (e.g. TCA cycle, amino acid biosynthesis, etc.), and their accumulation could be caused by changes in the activity of the associated cellular functionalities.

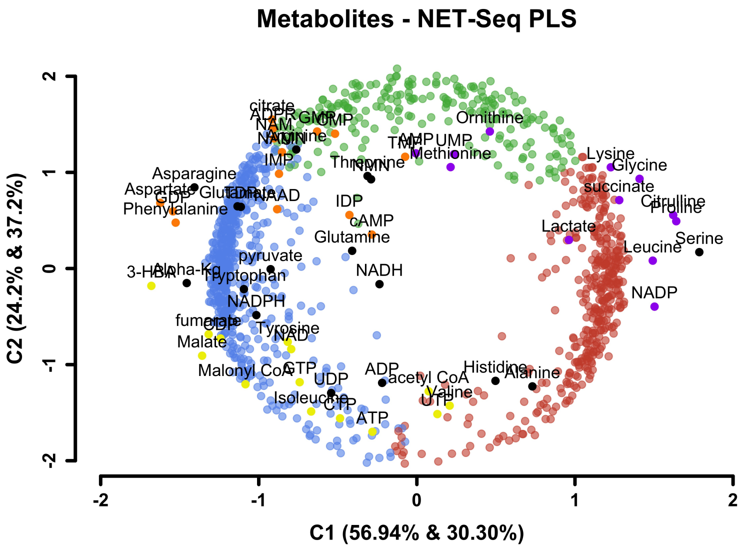


Figure 5.6: Biplot representation of the PLS model of metabolomics vs NET-seq data. Percentages are the explained variability of the RNA-seq (first) and metabolites (second) in each of the two components.

5.3.3.3 Chromatin conformation influences gene expression in a two-phase fashion

To study the impact of chromatin accessibility on gene expression during the YMC, we used PLS to model RNA-seq data as a function of ATAC-seq data. The resulting PLS model had an R^2 value of 0.65 and a Q^2 value of 0.25, which evidences a lower predictive power of ATAC-seq over gene expression compared to histone modifications and metabolites. Figure 5.7A shows the loadings of this PLS model, which revealed that ATAC-seq data discriminate gene expression into two phases (OX and RC) with the first component of the PLS model having a significant effect on the phase identity (ANOVA p-value < 0.05). These results corroborate that the ATAC-seq time-course data recapitulate a two-phase peak distribution, suggesting that chromatin occupancy is bi-phasic.

To further study the role of transcription factors on the YMC, we used MORE to model gene expression as a function of the expression of TFs identified by the ATAC-seq data. A specific gene-TF association is established when the ATAC-seq peaks contain the footprint of the TF located at the gene promoter region. These associations were used to establish the MORE models, and significant gene-TF associations were obtained when a given TF was detected as significant in the MORE model of a target gene. These significant gene-TF results were represented as a Gene Regulatory Network (GRN) (Figure 5.7B), where genes were linked to their significant regulators and colored by phase. The GRN highlights that OX and RC phases are indeed the most frequently regulated by TFs. Azf1 and Ime1 are the TFs that regulate the highest number of genes, and together with Rgt1 and Tda9 appeared to be the TFs that have the highest impact on RC genes, which is the most highly TF-regulated YMC phase.

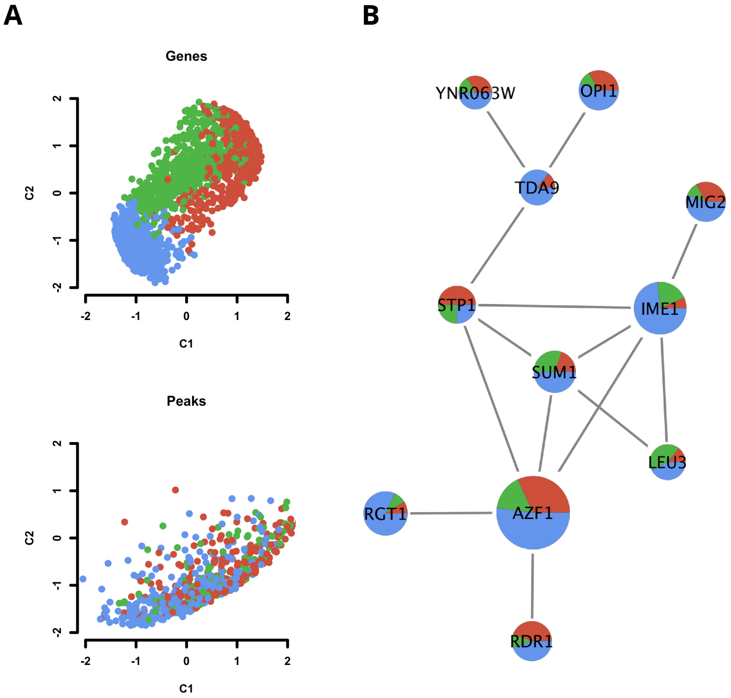


Figure 5.7: Modeling gene expression with ATAC-seq data. (A) The use of ATAC peaks to explain RNA variability results in a two-phase separation of RNA and ATAC features (ANOVA $<$ 0.05); (B) GRN of significant GLM relationships between TF expression and targets genes. A TF-gene target association was obtained from the relative position of ATAC-seq peaks in gene promoters.

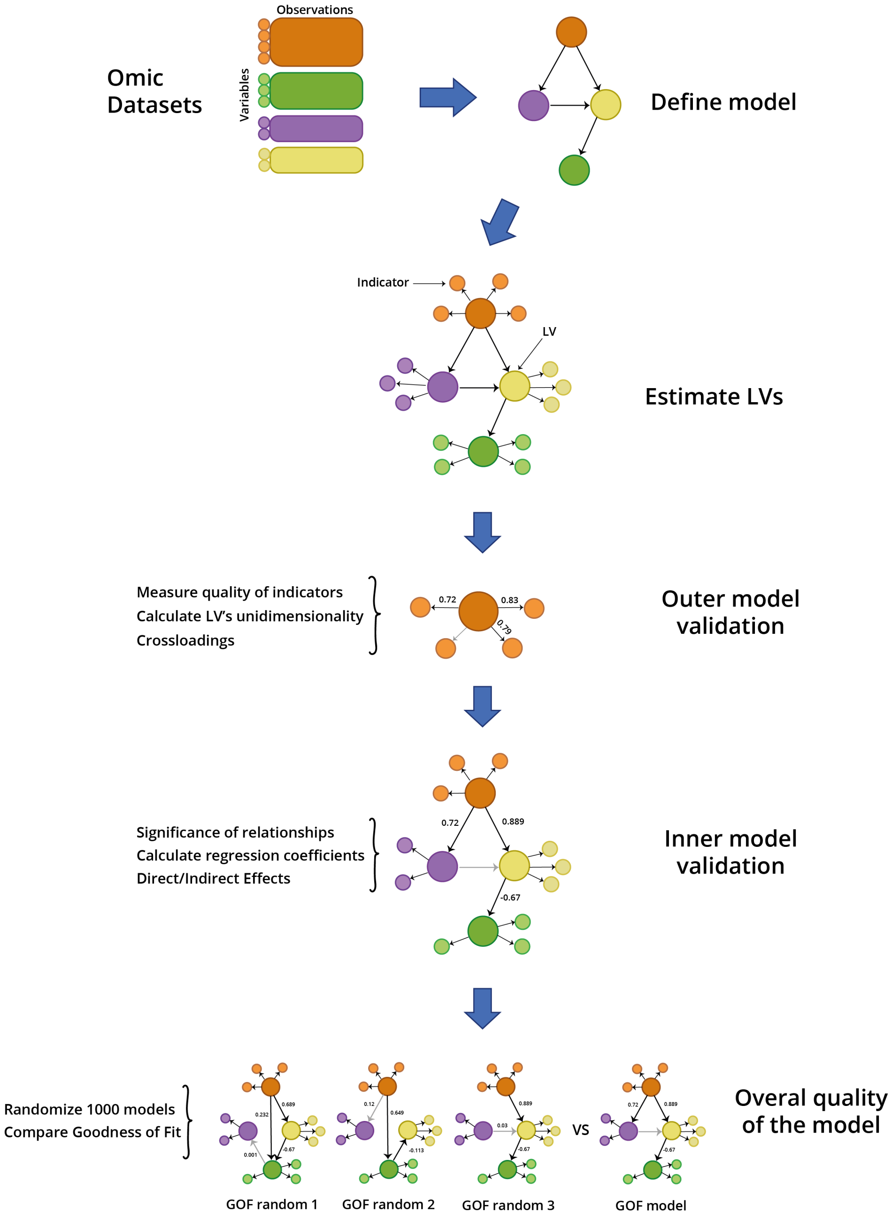


Figure 5.8: Schematic view of the steps to create a PLS-PM model.

5.3.4 PLS Path Modeling allowed the selection of metabolites that coordinate with epigenetic and gene expression changes

The main goal in our study was to determine how metabolic changes can drive gene expression through epigenetic modifications; although PLS and GLMs provided snapshots of the effects of different regulatory molecules on gene expression, these methods are limited in their representation of the succession of regulatory steps that are coordinated to establish the transcriptional regulation of the YMC. We applied PLS-Path Modeling (PLS-PM) to simultaneously model the relationships across all the studied omics and YMC phases. Notably, the omics layers and phases are not directly observed variables since we measure, for instance, the expression of a set of genes but not "gene expression in OX phase". One of the advantages of PLS-PM is that it can deal with unobserved variables (latent variables) that are defined by a set of indicators (observed variables, such as genes). Figure 5.8 shows a schematic view our implementation of PLS-PM for multi-omics modelling.

5.3.4.1 Model definition

The application of PLS-PM first requires that a causal model relating the latent variables (LVs) is defined. This causal model is the hypothetical network describing the relationship between the LVs considered. In our case, LVs were omic layers at different YMC phases, such as RC metabolites or OX gene expression. The network definition was based on our hypothesis about the relationship among regulatory layers, and PLS-PM tested whether the proposed network was supported by the available data. Therefore, we set metabolites to be the predictive variable of histone modifications, which is, in turn, predictive of nascent transcription (NET-seq), and this layer is predictive of TFs and gene expression separately (Figure 5.9).

Initial fitting of the model suggested that more than one variation pattern might be present for the metabolomics datasets and for the histone dataset at the RB phase (not shown) and, therefore, metabolomics was modelled as two LVs. Moreover, RB histone modifications were separated in H3K9ac and

H3K18ac, as these histone marks presented higher differences in their profiles in RB than in OX or RC.

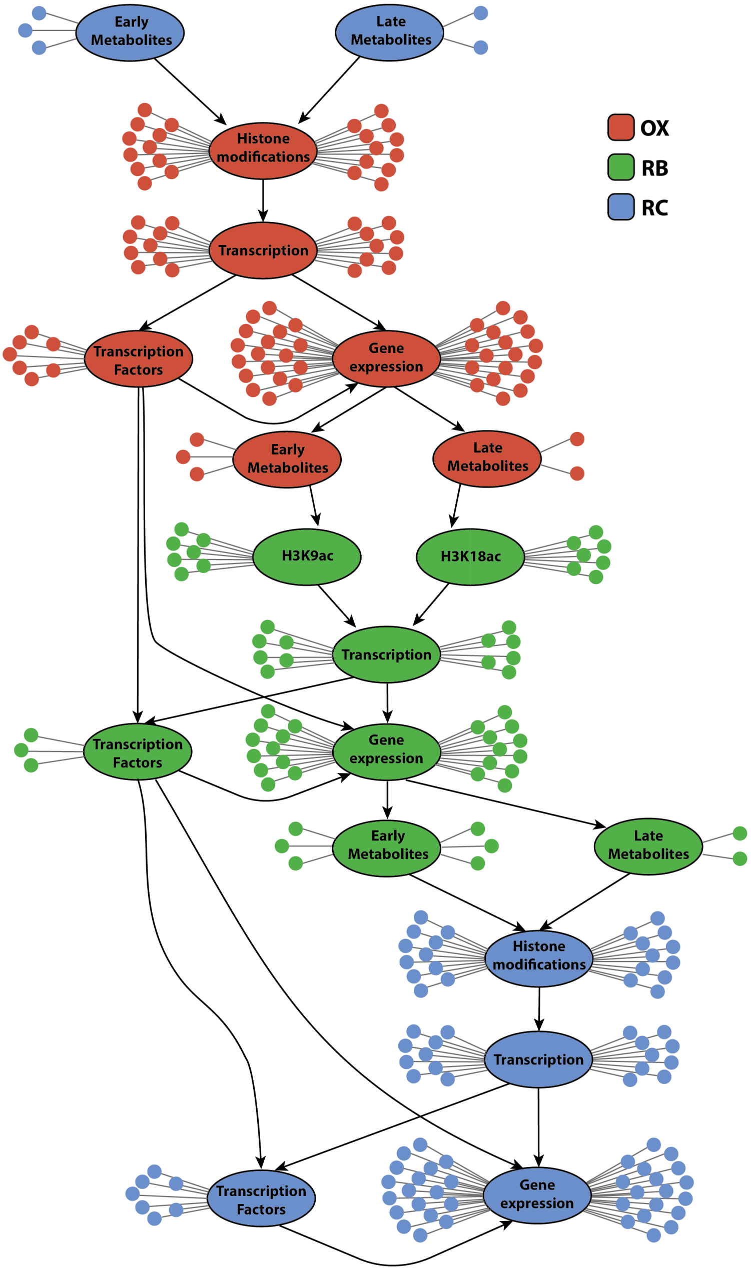


Figure 5.9: Inner model used in the PLS PM algorithm.

Table 5.3: Number of indicators present when defining the model, indicators whose sign had to be changed and final number of indicators per LV.

	Initial Indicators	Negative Indicators	Final Indicators
Early RC Metabolites	11	0	10
Late RC Metabolites	3	0	3
OX Histone modifications	955	0	691
OX Transcription	373	3	288
OX Transcription Factors	24	1	16
OX Gene Expression	1139	4	923
Early OX Metabolites	9	0	7
Late OX metabolites	3	0	3
RB H3K9ac	244	0	180
RB H3K18ac	88	0	51
RB Transcription	232	0	144
RB Transcription Factors	5	0	4
RB Gene Expression	657	0	514
RB Early Metabolites	11	0	9
RB Late Metabolites	3	0	3
RC Histone modifications	622	0	529
RC Transcription	564	0	453
RC Transcription Factors	30	0	25
RC Gene Expression	1238	0	1070

5.3.4.2 Validation of the outer model

When the inner and outer models are defined, and before estimating the strength of the relationships between LVs in the inner model, the outer model must be validated. Figure 5.10A shows an example of a LV and its indicators which, in this case, is the connection of the OX metabolites LV to its indicators, connected through a reflective relationship. The indicators with loadings lower than 0.7 were removed, as illustrated in Figure 5.10B. We changed the sign of indicators with a negative loading to ensure that all indicators in a LV were positively correlated. Table 5.3 summarizes the initial number of indicators, the indicators whose sign was changed and the final number of indicators in our PLS-PM model. In general, very few indicators needed a change of sign, and most of the initial variables were maintained as high-quality indicators of their respective LVs, which is consistent with the previous selection of maSigPro significant variables and their effective clustering in YMC functional phases.

Since all indicators in a LV should reflect the LV, they are expected to be highly correlated. This aspect is measured with unidimensionality, which captures whether the variability of the indicators follows similar trends. There are

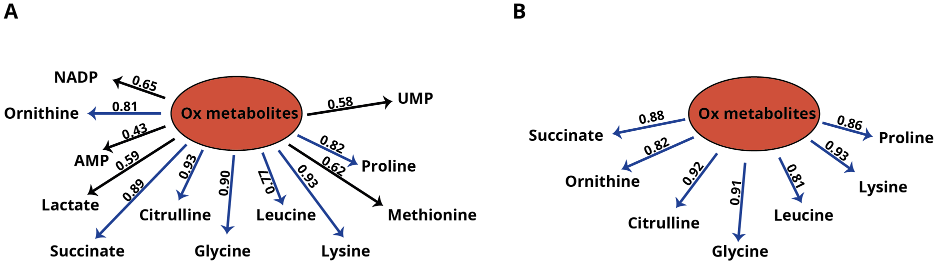


Figure 5.10: (A) Loadings of the indicators connected to the Latent Variable (LV) "Early Ox metabolites" with a reflective connection; some loadings are below the threshold of 0.7 and should be removed since they do not follow the LV values. (B) Corrected indicators for the "Ox metabolites" LV.

different ways of measuring unidimensionality. We used Chronbach's alpha (Eq 5.8) and Dillon-Goldstein's rho (Eq 5.9). Unidimensionality measures where calculated for the LVs in our model after correcting the low-contributing loadings and can be found in Table 5.4. All LVs presented a high unidimensionality (C-alpha $>$ 0.9; DG-rho $>$ 0.9; except for the LV RB Late metabolites where the two measurements were higher than 0.8). We concluded that the unidimensionality requirement was met in our model.

Finally, we computed cross-loadings and evaluated whether indicators have higher cross-loadings with other LVs than with the loading in their own LV. In our case, over 60% of the indicators had a higher loading than cross-loadings (Figure 5.11), which means that, as expected, they are more correlated with their own LV than with other LVs. For the remaining indicators, in 70% of the cases, the highest cross-loading corresponded to an LV in the same phase. In addition, most of them presented a difference smaller than 0.1 between the loading and the highest cross-loading. The remaining indicators (12% of the total set) belonged to genes with expression profiles that peaked early or late in their corresponding phase; since the loading with their corresponding LV was higher than 0.7, we kept them in the original construct.

5.3.4.3 Inner model validation

After the LV indicators were corrected and the outer model was validated, the coefficients of the inner model connecting LVs were calculated by means of re-

Table 5.4: Unidimensionality of the latent variables in the final model.

LV	Indicators	C alpha	DG Rho
Early RC metabolites	10	0.96	0.965
Late RC metabolites	3	0.79	0.878
OX histone modifications	691	0.999	0.999
OX Transcription	288	0.998	0.998
OX Transcription Factors	16	0.979	0.981
OX Gene Expression	923	0.999	0.999
Early OX metabolites	7	0.951	0.96
Late OX metabolites	3	0.916	0.947
RB H3K9ac	180	0.997	0.997
RB H3K18ac	51	0.992	0.992
RB Transcription	144	0.997	0.997
RB Transcription Factors	4	0.898	0.929
RB Gene Expression	514	0.999	0.999
RB Early metabolites	9	0.944	0.953
RB Late metabolites	3	0.817	0.893
RC histone modifications	529	0.999	0.999
RC Transcription	453	0.999	0.999
RC Transcription Factors	25	0.987	0.988
RC Gene Expression	1070	0.999	0.999

gression models. The quality of these regression models was assessed through their R^2 coefficient and their redundancy. Table 5.5 shows the values for these parameters, which indicates that LVs variability (except for Early OX metabolites and RB H3K9ac) was highly explained and predicted by their explanatory LVs. The lowest R^2 value was associated to Early OX metabolites, meaning that its variability was poorly explained by the OX genes LV, and other processes not considered in this analysis may be affecting Early OX metabolites.

The significance of the relationships connecting LVs is estimated by bootstrapping, which recalculates the model through 100 re-samplings of the indicators, and provides percentiles of the coefficient means (Table 5.6). Relationships with a non-significant coefficient were removed from the model.

The effect of the LVs on the other variables of the model was also estimated. Direct effects are those observed between two LVs having a direct relationship, and indirect effects refer to LVs connected through intermediate LVs. The magnitudes of these effects determine whether an LV has a greater effect through indirect relationships rather than direct. In our model, most direct interactions

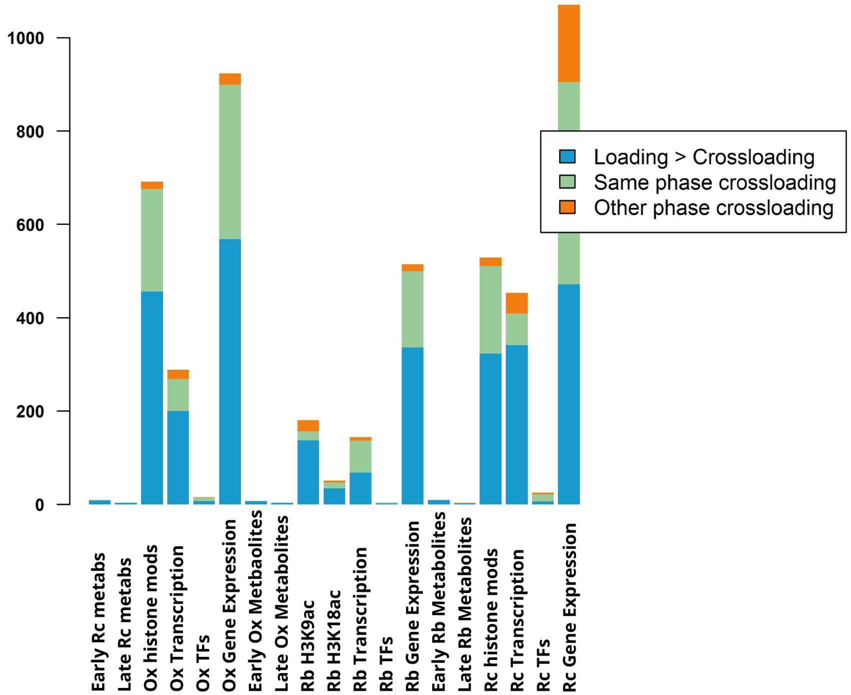


Figure 5.11: Cross-loading analysis of LV indicators. LV indicators are colored as a function of the cross-loading value.

had a stronger effect than indirect links (Figure 5.12), except for the relationship between RB Transcription and RB Gene Expression, which appears to be higher than RB Transcription to RB TFs. Nascent transcription (Transcription LV) and mature transcripts (Gene Expression LV) were expected to have a high correlation, so their relationship was initially tested in the model but was rejected. This rejection could have been caused by the different underlying data of indicators from Transcription and TFs. Since TFs measurements were obtained from the RNA-seq dataset, the correlation between these two LVs is expected to be higher than correlations with NET-seq. The indirect relationship between Transcription and Gene Expression is also strong in their direct interaction at OX and RC phases. The other strong, indirect effects are those that connect histone modifications with gene expression. As we have shown in the first chapter of this thesis, histone modifications act as regulators of gene expression; thus,

Table 5.5: Assessment of the inner model latent variables. These measurements assess the variance in the latent variables explained in the model.

LV	Type	R^2	Mean_Redundancy
Early RC metabolites	Exogenous	-	-
Late RC metabolites	Exogenous	-	-
OX Histone Modifications	Endogenous	0.829	0.613
OX Transcription	Endogenous	0.815	0.6
OX TFs	Endogenous	0.772	0.589
OX Gene Expression	Endogenous	0.982	0.766
Early OX metabolites	Endogenous	0.337	0.26
Late OX metabolites	Exogenous	-	-
RB H3K9ac	Endogenous	0.449	0.314
RB H3K18ac	Endogenous	0.851	0.611
RB Transcription	Endogenous	0.971	0.69
RB TFs	Endogenous	0.944	0.725
RB genes	Endogenous	0.963	0.75
Early RB metabolites	Endogenous	0.602	0.419
Late RB metabolites	Exogenous	-	-
RC histone modifications	Endogenous	0.774	0.582
RC Transcription	Endogenous	0.744	0.6
RC TFs	Endogenous	0.765	0.59
RC Gene Expression	Endogenous	0.979	0.794

their strong indirect effect is not surprising.

Table 5.6: Bootstrapping results of the inner model path calculation.

	Original	*Mean Boot*	*Std Error*	*perc 025*	*perc 975*
Early Rc Metabs ->OX Hist Mods	*-1.1018725*	*-0.9503112*	*0.18249866*	*-1.34395586*	*-0.62466393*
Late Rc Metabs ->OX Hist Mods	*0.9477295*	*0.7427699*	*0.22210517*	*0.16424103*	*1.0440692*
Ox Hist Mods ->OX Transcription	*0.9028036*	*0.9109843*	*0.02596369*	*0.85845487*	*0.96211353*
Ox Transcription ->OX TFs	*0.8787011*	*0.8865217*	*0.03078087*	*0.8272105*	*0.94656308*
Ox TFs ->OX Gene Expr.	*0.9911317*	*0.9877757*	*0.00898683*	*0.96211957*	*0.99601421*
Ox TFs ->RB TFs	*0.4408771*	*0.4283728*	*0.10688987*	*0.2357317*	*0.61003248*
Ox TFs ->Rb Gene Expr.	*-0.73999*	*-0.7254659*	*0.12770799*	*-0.9442748*	*-0.47754488*
Ox Gene Expr. ->Early OX Metabs	*0.5804976*	*0.6041701*	*0.0799758*	*0.4532748*	*0.77467371*
Early Ox Metabs ->RB H3K9ac	*-0.7103431*	*-0.7343058*	*0.14853708*	*-1.00174265*	*-0.46530786*
Ox.metabs1 ->RB H3K18	*0.9225372*	*0.9256713*	*0.01813135*	*0.88816599*	*0.95445327*
Late Ox Metabs ->RB H3K9ac	*0.5628542*	*0.5498135*	*0.244752*	*0.02439894*	*0.97844952*
Rb H3K9ac ->RB Transcription	*0.6468684*	*0.6367747*	*0.08703863*	*0.48081873*	*0.80501443*
Rb H3K18ac ->RB Transcription	*1.0471388*	*1.0630648*	*0.09751937*	*0.91314218*	*1.27904103*
Rb Transcription ->RB TFs	*0.8398549*	*0.8364671*	*0.04462586*	*0.74352342*	*0.91210932*
Rb.TFs ->RB Gene Expr.	*1.1036595*	*1.117202*	*0.08954498*	*0.9784086*	*1.31151756*
Rb TFs ->RB TFs	*-0.2617386*	*-0.2476937*	*0.10187633*	*-0.39644059*	*-0.02487273*
Rb Gene Expr ->Early RB Metabs	*0.7800899*	*0.7890915*	*0.0523488*	*0.66038985*	*0.86360979*
Late Rb Metabs ->RC Hist Mods	*0.8802137*	*0.886249*	*0.02685606*	*0.82719967*	*0.92390824*
Rc Hist Mods ->RC Transcription	*0.8623248*	*0.8706878*	*0.03146621*	*0.81376327*	*0.92312163*
Rc Transcription ->RC TFs	*0.7015443*	*0.7162244*	*0.07134182*	*0.60309929*	*0.87012146*
Rc TFs ->RC Gene Expr.	*0.9894294*	*0.989892*	*0.00283527*	*0.98385227*	*0.99510246*

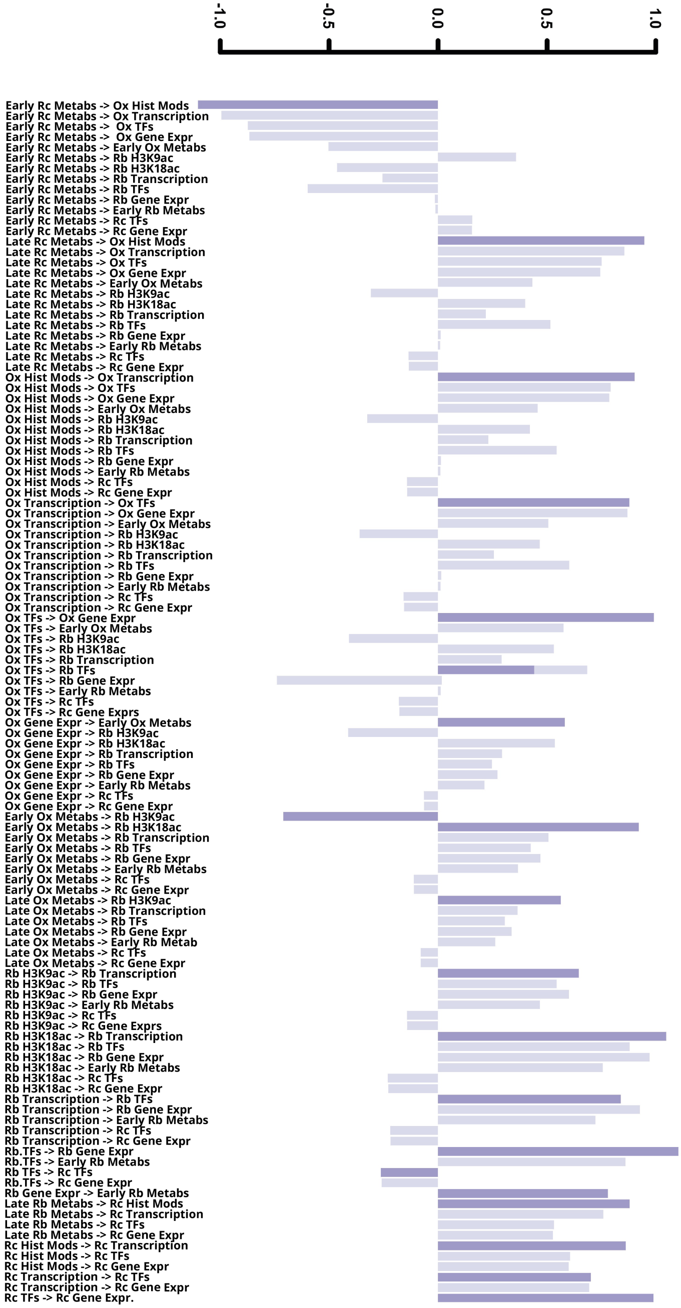


Figure 5.12: Direct and indirect effects of the Latent Variables in the PLS-PM.

5.3.4.4 Final Model

The final PLS-PM model estimated for our multi-omic dataset is presented in Figure 5.13. Eight relationships were found to be non-significant (gray arrows). Transcription relationship to Gene Expression was removed from the model due to the high correlation between Gene Expression and TFs, thus making impact of Transcription in the regression model insignificant. The flux from late metabolites to histone modifications had high and positive coefficients; this was also the case for the connections between histone modifications and Transcription, from Transcription and TFs and gene expression and early metabolites. Early metabolites showed significant connections with Gene Expression of the same phase, while late metabolites were usually predictors of the histone modifications of the next phase. This pattern was slightly different at the RB phase, were early OX metabolites were positively connected to RB H3K19ac; furthermore late OX metabolites had positive and significant coefficients explaining H3K18ac. These findings supports the previous analysis that suggested that there are significant differences between these two histone modifications in RB phase. Moreover, the relationship between early RB metabolites and RC histone modifications was rejected by the model, while the relationship between late RB metabolites and RC histone modifications was retained, suggesting a more direct impact of late RB metabolites in the epigenetic changes of RC.

The quality of the model is given by the Goodness-Of-Fit (GoF) parameter, which had a value of 0.778. However, as the connections among LVs were defined in the initial model, and most of them were supported by the data, one logical question is whether the significance of the model was biased by our assumptions. To assess this possible bias, we randomly generated 1000 alternative path models from the same set of LVs and tested whether our model had a higher GoF than the other random models. This analysis indicated that the GoF of our model was higher than 99.6% of the random models, further validating the significance of the results (Figure 5.14).

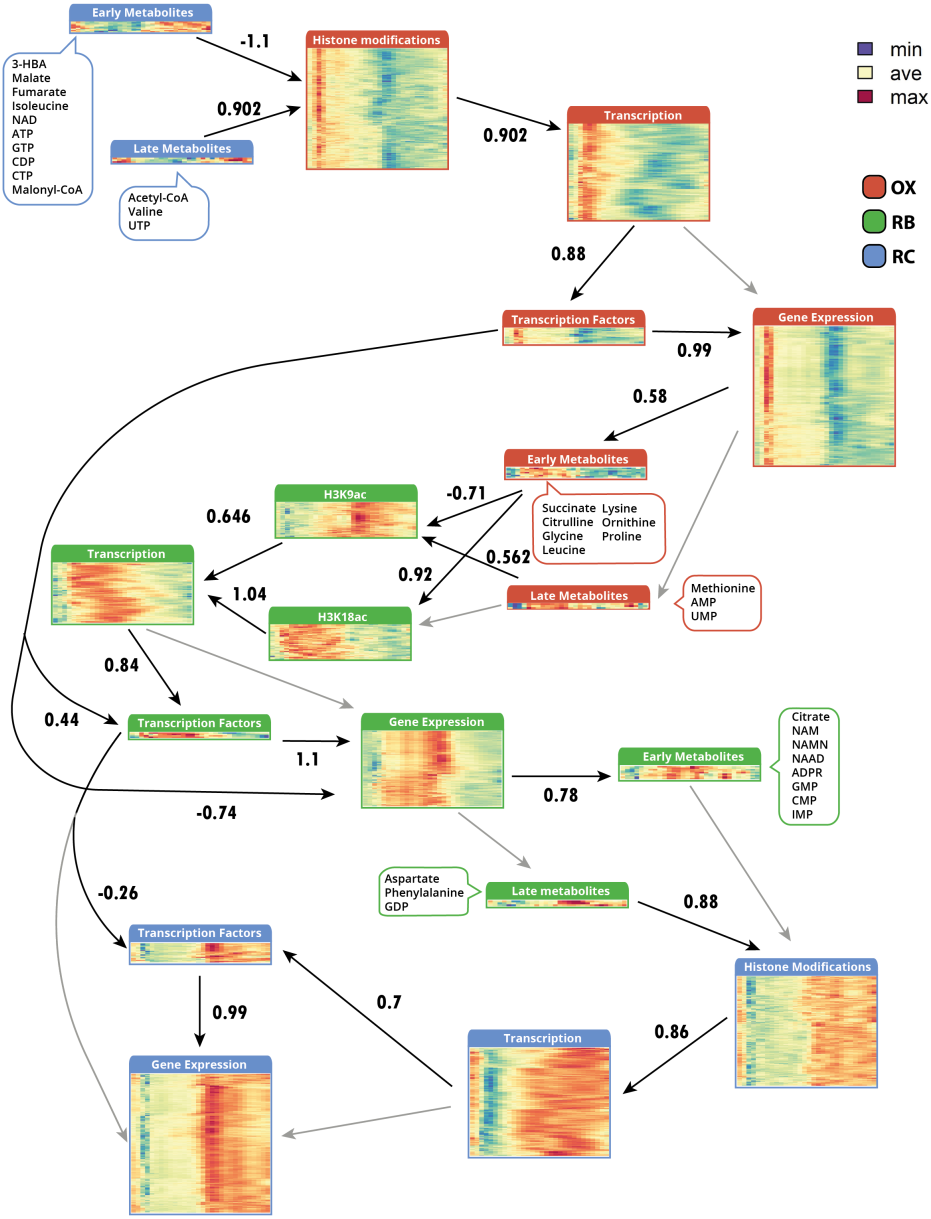


Figure 5.13: PLS-PM model of the YMC multi-omics dataset.

5.3.4.5 Functional interpretation

The goal of this study was to understand regulatory relationships between metabolites, epigenetic changes and transcription. We proposed a PLS-PM model that links metabolites with H3K9ac and H3K18ac measurements, which are related to transcription; transcription is, in turn, associated with transcription factors and gene expression. Overall, the relationships we proposed for the path model resulted in a highly predictive model, with higher GoF scores than most of the combinations tested in our analysis.

Generally, histone modifications were good predictors of transcription and, indirectly, of gene expression levels. However, there is a larger overlap between genes of histone modification indicators and gene expression indicators than with those of transcription indicators, probably reflecting the high signal-to-noise ratio characteristic of the NET-seq technology. This overlap, however, disappeared at the RB phase, where previous results suggested epigenetic regulation is reduced.

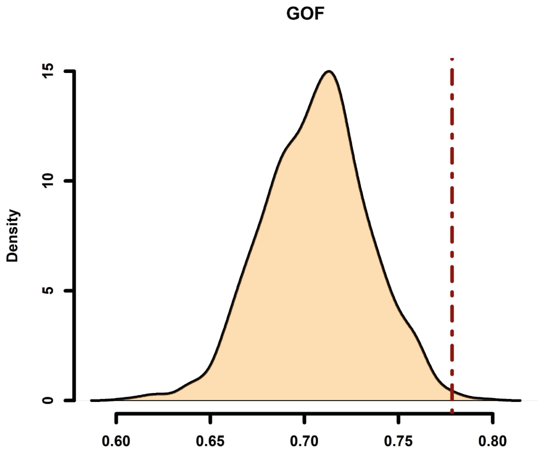


Figure 5.14: Goodness of Fit of 1000 randomized PLS-PM using the same LVs in all of them but randomizing their connections.

We found that late RC metabolites have a significant and positive explanatory capacity over OX histone changes, while the early metabolites of RC also contributeD to OX histone changes but with a negative sign. Early RC metabolites include NTPs, which are sensors of a high-energy cell state, as well as tricarboxylic acid cycle (TCA) intermediates fumarate and malate, Malonyl-CoA and NAD. Late RC metabolites included acetyl-CoA, valine and UTP. This means that the accumulation of acetyl-CoA, together with a depletion of compounds related to high-energy state, triggers the histone modifications observed in Ox phase. H3k9ac and H3K18ac affect genes mainly involved ribosomal assembly and translation. The overlap between the two histone marks is 70%, indicating that, to a great extent, these 2 histone marks act together in the regulation of their associated genes 5.15. The same type of enriched functionalities were found when assessing the impact of histone modifications in transcription and expression. The 16 TFs that are significant for this phase were involved in a variety of functions that include ribosome regulation, acetate production, sterol biosynthesis and glycolysis. OX phase resulted in accumulation of the amino acids leucine, proline and glycine in the early OX phase and methionine, AMP and UMP in the late OX phase.

The RB phase has a more complex and weaker regulatory pattern. First, the overlap of gene indicators at histone marks, transcription and expression is low, indicating a low synchronization pattern. RB H3K18ac is well explained by OX early metabolites, while OX late metabolites have a significant effect on H3K9ac changes. Moreover, only 6% of the indicators were shared between H3K9ac and H3K18ac 5.15, indicating that the two regulatory marks act pretty much independently in this phase. Moreover, most gene expression indicators overlap with H3K18ac, while only 12% of the gene expression indicators were shared with the H3K9ac indicators. RB H3K9ac showed a higher overlap with LVs from the RC phase, suggesting that it could act as a regulator of RC and have low to none effect in RB regulation. Additionally, the PLS-PM regression coefficient is higher from H3K18ac to RC transcription than for H3K9ac, supporting H3K18ac as the only histone mark with a mild impact in expression at RB. The indicators with the highest weight in the Gene expression LV corresponded to ATP synthesis, suggesting high energy production at the highest oxidative point of the cycle. The small fraction of RB's histone modification, NET-seq and Gene expression common indicators corresponded to genes with mitochondrial activity functionalities, suggesting that the chromatin control of gene expression in this phase might be restricted to mitochondrial energy processes.

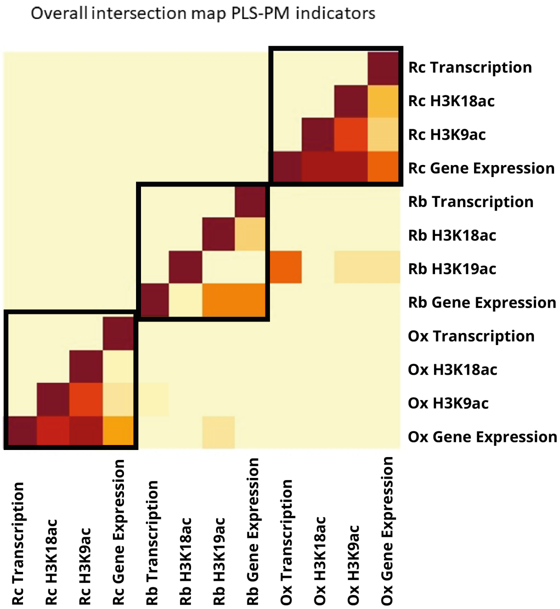


Figure 5.15: Intersection of the indicators of two LVs from the model. High intersection means that the indicators referred to the same genes in different omics measured.

RB phase early metabolites include NMPs and NAM derivatives, which are indicative of a low-energy state. NAM derivatives are repressors of sirtuin activity, preventing histone deacetylation, which could cause H3K18ac acetylation in this phase (Figure 4C). Late RB metabolites (Aspartate, Phenylalanine and GDP) are predictive of the Rc histone changes, although the mechanism of this potential regulation is not clear. Indicators of RC expression include vacuolar proteins and autophagosome complex components, while the RC genes with a coordinated change in histone modifications, transcription and expression are enriched in peroxisome activity, fatty acid degradation, eisosome activity, ER proteins and glucose-sensing MAPK signaling. These energy consumption processes result in accumulation of acetyl-CoA, malonyl-CoA or NTPs in RC metabolites, which triggers the start of OX phase as previously discussed.

5.4 Discussion

Biological rhythms are shaped by the combination of oscillations of multiple molecular components of the cell, including metabolites, epigenetic modifications and transcriptional activity [172]. Multiple studies have aimed to characterize the regulatory mechanisms that connect the molecular layers of biological clocks and to identify triggers that drive their oscillatory patterns. The study of the metabolic regulation of biological clocks has gained attention following breakthrough discoveries demonstrating that accumulation of certain metabolites results in changes in chromatin state [21, 198]. Metabolic oscillations can affect epigenetic modifications, such as histone acetylation/deacetylation, which in turn has an impact on gene regulation. In addition, gene expression closes this feedback loop by resulting in synthesis of enzymes that modify metabolite concentrations. Currently, there is a lack of methods to model such complex regulatory networks that are capable of capturing the dynamics of the oscillatory changes and describing the coordination among molecular layers. In this study, we used multivariate statistics, based on the analysis of the latent space, to develop an analysis strategy by which to characterize the molecular interconnections that operate in biological rhythms. We applied this strategy to the Yeast Metabolic Cycle (YMC), a regulated oscillatory system for which we obtained a comprehensive multi-omics dataset.

The YMC has traditionally been divided into three phases according to gene expression patterns. These phases are Oxidative (OX), Reductive Building (RB) and Reductive Charging (RC) [102, 187], although more recent studies have proposed that the metabolic oscillations of the cycle should be divided into two major rhythmic phases: High Oxygen Consumption (HOC) and Low Oxygen Consumption (LOC) [51, 126]. Our analysis of the variation patterns in different -omics datasets from the YMC suggested that the three-phase model system could accommodate most of the omics data, although the RB phase did not have such a strong and uniform behavior across different data types as the other phases did. The RB phase was clearly identified in the gene expression data, a result which indicated that regulatory mechanisms not captured by our analysis might be responsible for controlling gene expression at the transcript level in RB phase and that the regulation of RB phase does not have a strong epigenetic component. Recent research indeed showed that RB phase genes do not cycle at nascent transcription (NET-seq) and supports that this phase is post-transcriptionally regulated [51]. However, we did find a 3-phased pattern for metabolite oscillations, which might reflect the changes in enzymatic activity driven by gene expression.

PLS models relating two different omics data modalities added further evidence to the hypothesis that the RB phase is not transcriptionally regulated, as ATAC-seq PLS and GLM models did not find significant signals for chromatin accessibly at this phase. The PLS model relating gene expression variability with metabolomic changes (Figure 5.5) revealed the position of the cycle at which each metabolite accumulated. An interesting result of this analysis is that the lowest energy point of the cycle, where NAM derivates and NMPs accumulated, occurred after the HOC phase (the point of highest energy production); this suggested that the system could be using energy produced by other cellular processes, such as cell division, that is known to start at this stage of the cycle [20]. NAM derivatives are known sirtuin inhibitors [12], and their accumulation at early RC phase could promote histone acetylation of genes in RC phase, wherein most H3K18ac takes place according to MORE analysis 5.4. Malonyl-CoA and acetyl-CoA accumulated at the beginning of OX phase, the stage of the YMC at which previous studies reported their influence in the epigenetic regulation of growth genes [187].

The PLS-PM analysis of the multi-omics data allowed us to propose a regulatory network that connects metabolites with gene expression through the activity of gene expression regulators (Figure 5.13). OX phase regulatory metabolites included the NTPs, Malonyl-CoA and TCA intermediates. Acetyl-CoA was present in the late OX metabolites LV, suggesting a direct role in the regulation of OX genes, as has been reported in previous studies. Interestingly, OX genes include multiple growth genes (KSS1, SHO1); furthermore, previous studies have reported that acetyl-CoA accumulation leads to the epigenetic activation of growth genes. The NAD and NTPs are cofactors of the lysine producing enzymes, and lysine production genes were among the indicators with highest weight in OX histone modifications, NET-seq and gene expression, which also included other amino acid biosynthesis functionalities. Furthermore, the accumulation of ATP at this point of the cycle could drive INO80 activity, an ATP-dependent chromatin remodeler that has been reported to activate mTOR pathway in OX phase to regulate amino-acid biosynthesis [61].

The accumulation of amino acids at the end of OX phase was caused by the high activity of amino-acid biosynthesis enzymes, as described for Lysine. Although RB LVs captured some of the variability in this phase, very few any genes shared indicators with NET and histone modification LVs indicators, corroborating previous conclusions on the limited transcriptional regulation at the RB phase, at least for the -omics layers studied here. It is possible that low epigenetic regulation is related to the start of cell cycle in this phase, thus posing a limited chromatin accessibility that could explain the low resolution of ATAC-seq sampling at this stage of the cycle. The genes found in this phase indicators were mostly related to mitochondrial activity and ATP production. Although most TCA metabolites' indicators accumulated at RC metabolites LV, succinate (an inhibitor of TET demethylases) accumulation at the beginning of RB and methionine (methyl donor for DNA/histone methylation) accumulation in mid-RB suggested an active role of methylation at this stage of the cycle; however, no methylation marks were included in our model as the previous analysis described in Chapter 3 showed that of H3K4me3 was not relevant to gene expression regulation in the YMC. Figure 5.2 shows that H3K9ac has a late accumulation at this phase, which could explain its low coefficient in its relationship with transcription; it is even possible that this histone mark in RB phase is more associated with RC phase, as suggested by its profile.

Although ATP production was among the most extended functionalities in RB genes LV, RB's early accumulation of metabolites was characterized by NMPs and NAD-derivatives, suggesting that produced ATP was consumed in other cell processes. Among the RB functionalities, we could also distinguish

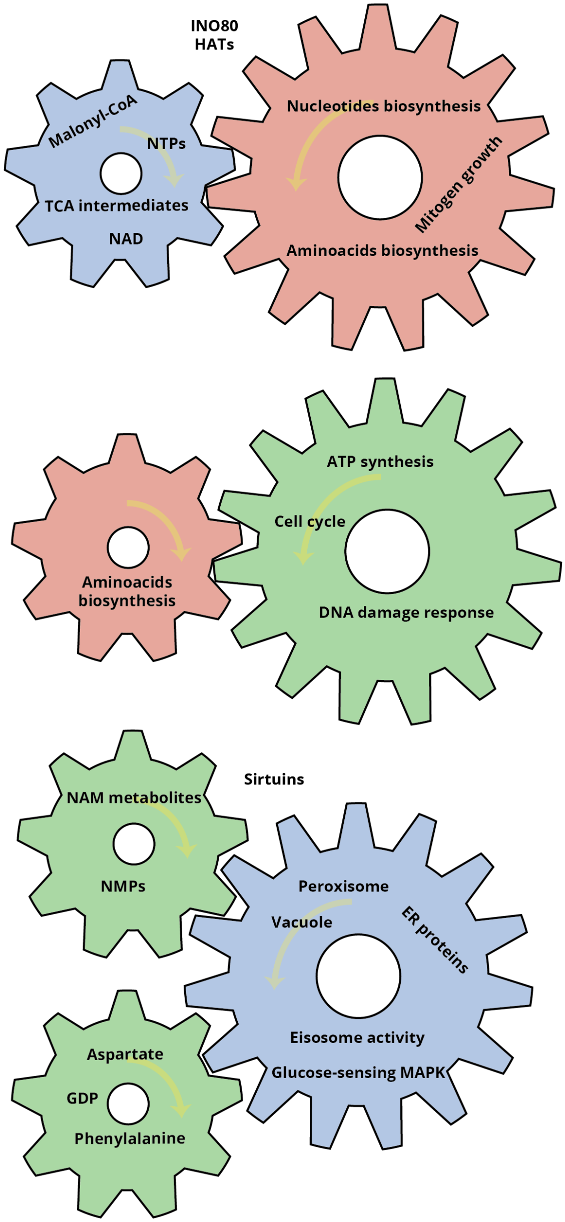


Figure 5.16: Functional characterization of the interplay between metabolites and gene expression. HATs, INO80 and Sirtuins act as epigenetic remodelers that signal gene translation by promoting histone acetylation (HATs) and activating aminoacid biosynthesis (INO80) in OX phase and inhibiting sirtuins to promote acetylation in RC Phase (NAM-mediated sirtuin inhibition). RB phase does not present any known epigenetic signaling.

cell cycle genes, which led us to hypothesize that ATP was consumed in such an energy-demanding process. NAM accumulation in the late RB metabolites phase suggested an indirect regulation of acetylation driven by repression of sirtuins activity. Aspartate, Phenylalanine and GDP accumulated in Late RB metabolites and could have a key role in the epigenetic regulation of RC phase in the YMC. Aspartate has not been directly linked with epigenetic changes but has been indirectly linked with methylation through transmembrane transport of alpha-ketoglutarate [120]. As it has been discussed previously, methylation was not included among the histone marks studied in the PLS-PM model; therefore, further study is necessary to assess the impact of aspartate in the YMC. Finally, our model indicated that RC phase is transcriptionally controlled by H3K18ac and by fatty-acid and glucose responsive TFs, leading to the activation of peroxisome, vacuole and fatty-acid beta oxidation functions and to the accumulation of NTPs, malonyl-CoA and acetyl-CoA, which in turn trigger the chromatin changes at OX phase as previously discussed.

To conclude, in this work we presented a set of multivariate methodologies for the integration of multi-omics datasets. We have applied this analytical framework to the YMC, as this system presents a strong coordination across multiple molecular layers and is therefore suited for the assumptions of the PLS-PM methodology. Although this analytical strategy could be potentially applied to any multi-omics dataset, the analysis presented here benefits from the highly coordinated and oscillatory nature of biological signals, which are easy to capture by factor analysis methods. In this way, we were able to identify the metabolites that correlate with changes in epigenetic marks, and which potentially represent compounds that drive chromatin modifications affecting gene expression controlling the progress of the YMC. ATP and NAM-derivatives emerge as major drivers of OX and RC phases together with acetyl-CoA. Other metabolites, such as Aspartate, show interesting associations with H3K9ac, H3K18ac and gene expression in RC phase, although the mechanisms that could mediate these associations are still unknown.

Chapter 6

A computational pipeline for Pathway Reconstruction on the Fly

6.1 Introduction

Pathway databases are important bioinformatics tools that support genomics research by providing useful representations of molecular processes. Biological pathways interconnect proteins and metabolites and represent reactions that are necessary to enable cellular functioning. There are two main types of biological pathways: metabolic pathways and signaling pathways. Metabolic pathways consist of a chain of enzymatic reactions where enzymes transform metabolic substrates into products to synthesize the basic building blocks of cellular architecture and manage energy fluxes. Signaling pathways are the processes by which chemical signals are transmitted through the cell, which ultimately results in a cellular response.

There are multiple bioinformatics resources that contain curated pathways: KEGG [85], Reactome [34] or MetaCyc [87] provide mechanistic representations of metabolic and signaling reactions, and are widely used in genomics research to characterize biological processes. STRING [174] and Omnipath [189] feature networks that represent protein-protein interactions and signaling networks, respectively, and PAZAR [146] presents gene transcriptional regulation data. Pathway databases often make use of computational resources such as orthology analysis, text mining and high-throughput assays to support their pathway models, which are usually manually curated by knowledge-domain experts who verify content and add or remove elements prior to public release.

Manual curation of pathways is long and tedious, and although curated pathways offer reliable, well-structured information, their “create-and-release” nature implies several constraints. First, while cellular processes consist of a virtually unlimited set of interconnected reactions, pathways select and pack them as functional blocks, with their topology and boundaries defined by the curator’s choice. Users are therefore restricted to the provided pathway view, which may not be the best way to represent a domain of interest, as the database may include non-relevant information or have boundaries that exclude the reactions of interest. Second, biological research is highly dynamic, and novel scientific discoveries published in the literature may take years to consolidate in the curated, static pathway databases. This means that investigators at the forefront of their research may have difficulties in finding new or updated pathways in established databases, thereby missing the opportunity to use pathway analyses to support their studies.

The scientific community is aware of such limitations, and different solutions have been proposed to offer dynamic pathways that adapt to the needs of the researcher. For example, Reactome [34], Biochem4j [173] and LitPathExplorer [171] all allow a versatile interplay between the user and the database content, while Wikipathways [91] is a community resource that allows for a faster incorporation of novel discoveries. Unfortunately, interconnected and community-maintained databases still fall short in providing tailored biological pathways for new research fields. For example, in public pathway databases, we failed to find an integrated representation of the connection between carbon metabolism and epigenetic histone modifications leading to control of gene expression, a research field of growing interest for which a wealth of scientific literature is already available [21, 129, 198]. There is also a lack of pathway databases that offer reliable information on non-model organisms, which have a clear under-representation in current static resources.

Although pathway databases may take time to consolidate novel scientific knowledge into the pathways structures, this knowledge can be accessed from the literature source. BioNLP, or Biomedical Natural Language Processing, brings together methods that extract information from texts in the biomedical and molecular biology domains. Biomedical text mining has gained ground in recent years and a large variety of Natural Language Processing (NLP) tools are now available that analyze scientific manuscripts to extract meaningful information. The BioNLP shared task (BioNLP-ST) is a text mining competition that releases yearly data annotations to attract groups that develop text mining engines for the extraction of molecular biology information from the literature [95]. Annotated data is necessary for the usage of text-mining engines, as they provide the training dataset for the machine learning text-mining algorithms. Current annotated datasets include different knowledge domains such as GE (Genia Event

Extraction), CG (Cancer genetics) or GRN (Gene Regulation Network) among others. Although BioNLP tools extract biological relationships and phenotype associations from online papers, there are no tools today that use these technologies to provide dynamic construction of novel pathways. Such tools would offer the versatility currently required in a fast-evolving research environment and in a user-dependent manner.

In this work we present Padhoc, a pipeline for Pathway reconstruction ”on the fly”. Padhoc combines text-mining BioNLP resources with curated databases and Neo4j’s strong visualization capabilities [130] to create novel, up-to-date, biological pathways tailored to the specific needs of users. We demonstrate Padhoc’s performance on a set of well-known pathways and illustrate its potential for research in model and non-model species by creating two pathways: the link between metabolism and epigenetic control of gene expression in Homo sapiens, and the biotic stress response in Citrus sinensis. Padhoc is available at https://github.com/ConesaLab/padhoc.

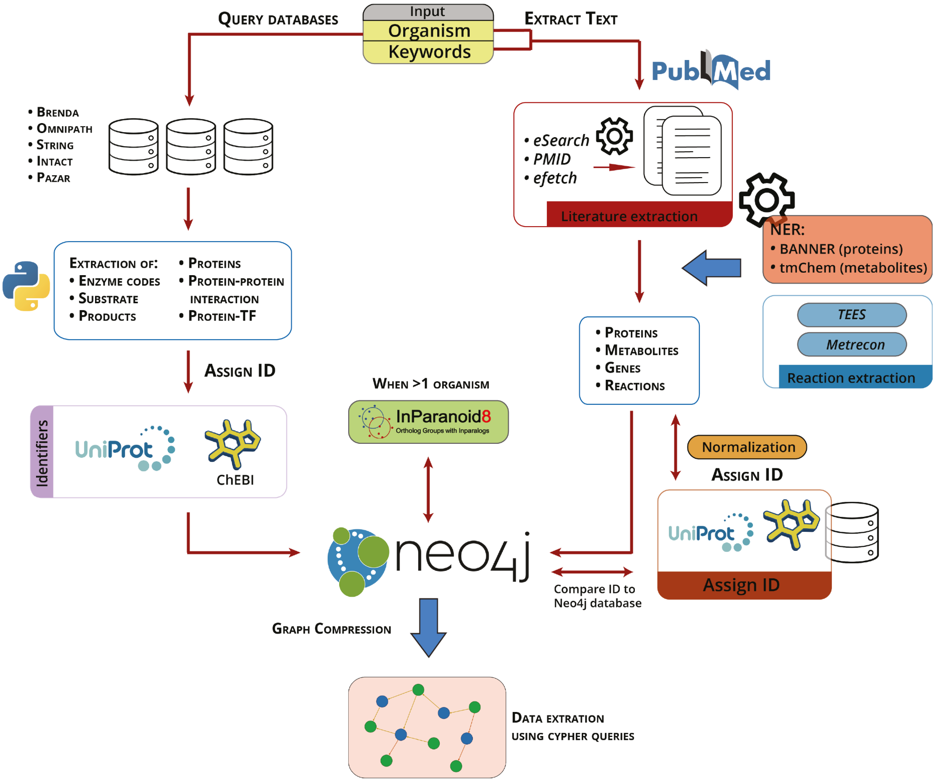


Figure 6.1: Computational pipeline implemented in Padhoc. The organism name keyword (Left) allows access to different public databases to extract genes, proteins, compounds and molecular relationships. Features are assigned UniProt/ChEBI IDs and incorporated into Neo4j. Organism and pathway keywords (Right) are used to recover relevant literature from Pubmed. Genes, proteins, compounds and relationships are extracted using NER resources, and the text is normalized and incorporated into the Neo4j database after assigning UniProt/ChEBI IDs. InParanoid is used to support homology-based pathway modelling. New pathways are visualized using Neo4j graphical resources.

6.2 Methods

6.2.1 Databases and software

Padhoc uses a number of public databases, information resources and functions to construct pathways on the fly. Databases include Brenda [162], Omnipath [189], IntAct [92], String [174], Pazar [146] and ENCODE transcription factor data [50]. PubMed is the source of literature data. Functions and software utilized to extract information from these resources are detailed in Table 6.1.

6.2.2 Text-mining resources

6.2.2.1 Named entity recognition

Named Entity Recognition (NER) tools identify and classify named entities that appear in unstructured text. Padhoc makes use of NER systems to extract names from selected literature, classify them into protein or metabolite names and identify reactions in the text (Figure 6.2, top). The NER systems incorporated in Padhoc are BANNER, for protein recognition, and tmChem, for metabolic recognition. These two systems have been described as gold standards in biomedical text extraction and both have shown high Precision, Recall and F-Measure scores at the BioCreative Challenge Evaluation [100].

Table 6.1: Software used for Padhoc's architecture, including algorithms used, input data, recovered output and source for software utilization.

Software	Function	Input	Output	Activity
BANNER	banner	Sentences, corpus	XML entity recognition	Extract protein and metabolite mentions from text, stores and classifies them in XML formated file
tmChem	tmchem	Sentences, corpus	XML metabolite recognition	Extracts metabolite mentions from text and reports them with ChEBI identifier in XML formated file
TEES	tees	Sentences, corpus	XML relationship	Extracts relationships from text sentences between banner/tmChem-recognized entities
Metrecon	metrecon	Sentences, corpus	XML relationship	Extracts metabolic relationships from text sentences between banner/tmChem-recognized entities
ChEBI	chebi	metabolite name	CHEBI Identifier	Assigns a Chebi identifier to a metabolite name
UniProt	uniprot_request (bioservices)	protein name	UniProt Identifier	Assigns an UniProt identifier to a protein name
BRENDA	getSubstrate	Enzyme Code	List of substrates	Extracts substrates of specified reaction
BRENDA	getProduct	Enzyme Code	List of products	Extracts products of specified reaction
BRENDA	getSynonyms	Enzyme Code	List of synonyms	Extracts synonyms of the Enzyme of the reaction
BRENDA	getRecommendedName	Enzyme Code	Recommended name	Extracts main name of the Enzyme of the reaction
Omnipath	In-house scripts	Specie	PTMs in CSV	Extract signaling relationships from source database
STRING	In-house scripts	Specie	Interactions in CSV	Obtain Protein protein interaction relationships from source database
PAZAR	In-house scripts	Specie	Regulation ins CSV	Extract transcription factor-gene regulation relationships from source database
InTACT	In-house scripts	Specie	Interactions in CSV	Obtain Protein protein interaction relationships from source database
Neo4j	GraphDatabase.driver	user,password	Neo4j connection	Connection to Neo4j graph database
Neo4j	driver.session.run	cypher query	Interaction with Neo4j	Runs cypher queries to modify/extract information from Neo4j Graph Database
InParanoid	inparanoid.pl	fasta files with protein sequences	XML of orthologous relationships	Detects orthology by running homology matches between protein sequences

6.2.2.2 Relationship extraction

Turku Event Extraction System (TEES) is a natural language processing system used for the extraction of events and relationships from biomedical text [13]. TEES was developed to extract molecular interactions from biomedical literature and has been used in multiple BioNLP shared tasks [95]. Metrecon is a text-mining system derived from TEES, which is optimized for the identification of metabolic reactions [142]. Both TEES and Metrecon use NER systems to identify proteins and metabolites in scientific literature, which are then used as a baseline for extraction and classification of their molecular relationships described in the text (Figure 6.2, bottom). TEES and Metrecon were originally developed to work with BANNER and in this work, we have modified their structure to also use tmChem for metabolic text recognition. Padhoc's implementation of these NERs facilitates the selection of one or multiple corpora and methods to be used for training, ensuring that all types of biomolecules are efficiently extracted.

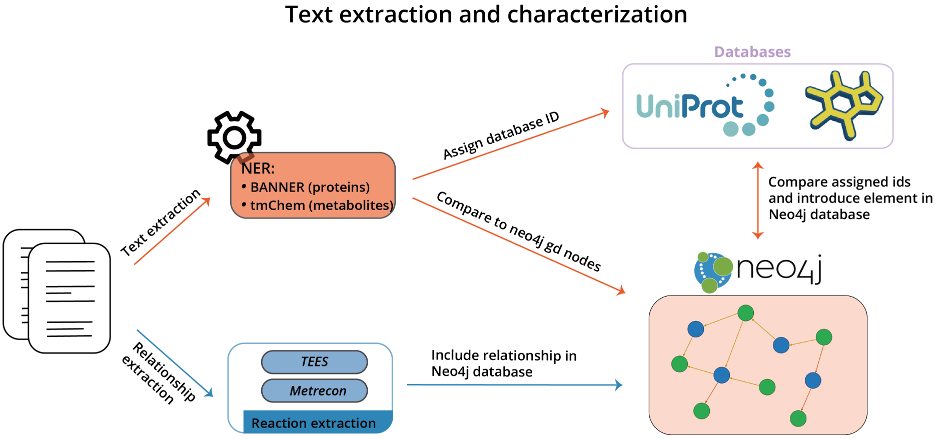

Text extraction and characterization
Databases
UniProt
NER:
• BANNER (proteins)
• tmChem (metabolites)
Text extraction
Assign database ID
Compare to neo4j gd nodes
Compare assigned ids and introduce element in Neo4j database
neo4j
Relationship extraction
TEES
Metrecon
Reaction extraction
Include relationship in Neo4j database

Figure 6.2: Extraction of text from literature is approached using BANNER and tmChem, which are Named Entity Recognition (NER) systems that extract entities from text (top). Entities extracted from text must be normalized to assign a UniProt/ChEBI ID. Relationships are extracted from text using TEES and Metrecon NLP software (bottom).

6.2.3 Neo4j graph database

Neo4j is a graph database that has been widely used to store connected data [130]. Graph databases store information as nodes and edges, which represent relationships between the nodes. Both data types contain properties, where additional information is stored. Graph databases allow easy and fast retrieval of information, which, combined with the graph structure and strong visualization capabilities, makes them the perfect environment to store biological networks. Neo4j grants the extraction of connected information (Figure 6.3A) and offers engines for manual adjustments of the network (Figure 6.3B).

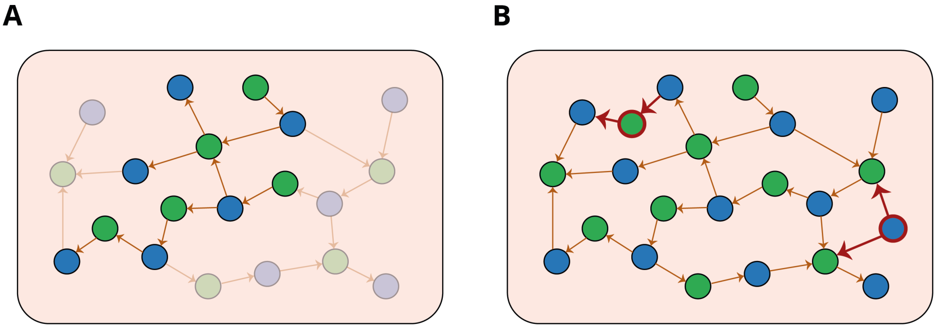


Figure 6.3: (A) Neo4j extracts subgraphs from the information stored in the graph database using cypher queries. (B) Red nodes represent elements manually included in the graph database.

6.2.4 Add pathway databases to Neo4j

Information retrieved from the pathway databases listed in previous section is the first source of information included in Neo4j; protein nodes are labelled either as “Proteins” or “Enzymes”, and metabolites are labelled as “Compounds”. Protein information used to fill node properties is obtained using UniProt API SOAP access [32], compound information is obtained through ChEBI API [36]. The list of properties extracted for each element is listed in Table 6.2. Edges connect database nodes that have a connection in any of the databases, and the edge is labelled with the name of database where the connection is obtained. Properties added to the edges vary depending on the nature of the relationship. Edges labels and properties are listed in Table 6.2.

Table 6.2: Node types and properties stored by Padhoc in a Neo4j database.

Node Label	Properties	Description
Compound	ID, chebiID, compoundName, textname, sentences, PMID_Tnb	Compound node
Protein	id, uniProtEntryName, uniprotGeneName, uniprotID, uniprotProteinName, specie, textname, PMID_Tnb, sentences	Protein node, also enzyme nodes if they catalyze a reaction
Enzyme	id, uniProtEntryName, uniprotID, specie, textname, sentences, PMID_Tnb, ECs, synonyms	Enzyme node
Compressed	id, uniprotIDs, Ecs, CompoundNames, chebiIDs, sentences	Compressed nodes, clusters of molecules

6.2.5 Add text information in Neo4j Graph Database

Once proteins and metabolites from text are extracted and classified using NER systems, entities must be characterized prior their incorporation into the graph database. Entities are referred in the text with a variety of names and, in order to homogenize the mentions to the same entity, names must be normalized and assigned a UniProt ID or ChEBI ID if they are proteins or metabolites, respectively (explained in detail in next section). After the text has been normalized, there are two types of text entities: those that were assigned an ID (ChEBI or UniProt depending on the nature of the molecule) and those where the text could not be assigned an ID. Entities that have been assigned an ID are searched for in the Neo4j database to check whether the entity is already present in the graph database. If it is not present, a new node is created; if the entity is present, only the text information is added to the already existing node (Table 6.2).

Relationships extracted from the NER systems are also included in the Neo4j database. If two database elements are found to be connected in the text, this connection is transformed into an edge that links both nodes in the graph database if they were not previously connected using other databases information. The edges derived from text mining are filled with properties that include types of relationships identified in text and number of appearances of the relationship in text, among others (Table 6.3).

Table 6.3: Type of relationships extracted by Padhoc and stored in Neo4j database for pathway reconstruction.

Label	Properties	Type
Brenda_database	ECs, reactionsBrenda, species	Enzymatic reactions
TM_relationship	corpora, nbs, query, reactionTypes, sentences, species	Text relationship
Omnipath_database	ptm, species	Signaling
PAZAR_database	metabod, pmid, species	TF-gene regulation
String_database	No properties	Protein-protein interaction
Intact_database	detection_method, experimental_score, interaction_method, interaction_type, species	Protein-protein interaction
Compressed_relationship	Sentences	Condensed graph
Compressed_to	No properties	Condensed to uncondensed
Orthologous_relationship	No properties	Orthology between proteins

6.2.6 Entity normalization

Names extracted from text-mining are highly variable in their notations and require normalization prior to assigning a UniProt/ChEBI identifier. We implemented a normalization procedure in Padhoc that follows the NACTEM parsing guidelines [185], which includes conversion to lower case, deleting spaces and removing isomer tags, among other rules (see full list at Table 6.4). After normalization, text entities are matched to existing Neo4j nodes using the difflib Python library, a text similarity searching tool that finds existing nodes that have a similar node name as the text entity. Text entities are then assigned UniProt/CheBi IDs combining the extracted Neo4j ID with the ID assigned by UniProt knowledgebase or tmChem.

Table 6.4: List of approaches used for normalization of the text name extractions.

Normalization rules
Lower case
Delete spaces
Delete characters: &, ;, and brackets
Delete numbers from beginning of string
Delete steroisomerity: l, d, r, s
Delete EC, alpha- and beta- at the begining of the string
Delete words gene and protein from string
Delete words beta, gamma, atom and complex
Delete acid and ate termination
Delete terminations ine, ase, ose

6.2.7 Homology

Padhoc enables pathway reconstruction for non-model species. When two organisms (i.e., a model and a non-model organism) are included in the keyword search, the Neo4j database is filled with information from public databases and text extracted from articles for both species. Proteins present in the graph database belong to the queried species and are searched for homology using their UniProt IDs on the InParanoid [139] web server (http://inparanoid.sbc.su.se/cgi-bin/gene_search.cgi). When orthologous relationships are not available in InParanoid, Padhoc extracts the protein sequences from the UniProt knowledgebase and performs pairwise Blastp similarity searches between the protein sequences from the species. Blastp results with a bit-score $>$ 40 are submitted to InParanoid v4.1 for orthology evaluation and are eventually incorporated into the Neo4j database (Figure 6.4). Proteins with inferred orthology relationships are connected in the database using an edge with the label 'Orthology_relationship'.

Neo4j's structure allows for the storage of the networks from the different species, which are connected at any of the orthologous nodes. This enables the use of network connections from one organism when the information is missing for the other organism, thereby avoiding pathway fragmentation in the non-model species.

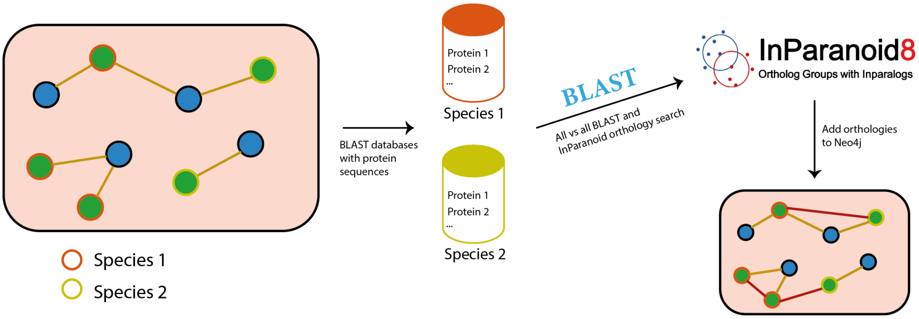


Figure 6.4: Padhoc can include more than one species in the Neo4j graph database and can search for orthologies using InParanoid. These orthology searches include orthologous information in the network that aid in the reconstruction of pathways from non-model species.

6.2.8 Graph compression algorithm

Database elements may still contain redundant information; for example, they can represent different isomers of the same compound or two orthologous proteins with the same gene name. In order to further compress data, semantic similarity matrices of protein and metabolite names are constructed using in-house scripts. Briefly, similarities are calculated using difflib after normalizing metabolite and protein names, and clusters of semantically similar features are obtained using the DBSCAN function from scikit-learn [101] (eps=1.0, min_samples=1). These clusters are included in the graph database as compressed nodes connected to their corresponding components (Figure 6.5). Singleton nodes are also included in the compressed graph.

6.2.9 Padhoc validation

Padhoc pathways were validated by reconstructing 13 well-established E.coli pathways (Table 6.5) and comparing them with their annotations in the MetaCyc database [87]. For this evaluation, Padhoc was directly fed with the scientific literature specified for each pathway in MetaCyc. Entities and relationships from the Padhoc-created pathways that occur in or are absent from the reference pathways were treated as true positives and true negatives, respectively, and sensitivity and specificity were calculated. Evaluation was performed by filtering entities for increasing levels (from 0 to 5) of literature support, i.e., number of times the entity appeared in the literature.

Additionally, Padhoc reconstruction of the *E. coli* Pantothenate and Coenzyme A biosynthesis pathway was manually evaluated by human assessment of the relevance of novel Padhoc pathway components and relationships. Features in the Padhoc Pantothenate and Co-A reconstructed pathway were given a relevance score from 1 to 3, where 1 means no relevance, 2 indicates inconclusive relationship, and 3 means high relevance for the pathway.

Finally, Padhoc was assessed for its ability to reconstruct two de novo pathways. The human histone acetylation pathway was constructed using the keywords ”Homo sapiens” and ”histone acetylation”, and the stress response in cit-

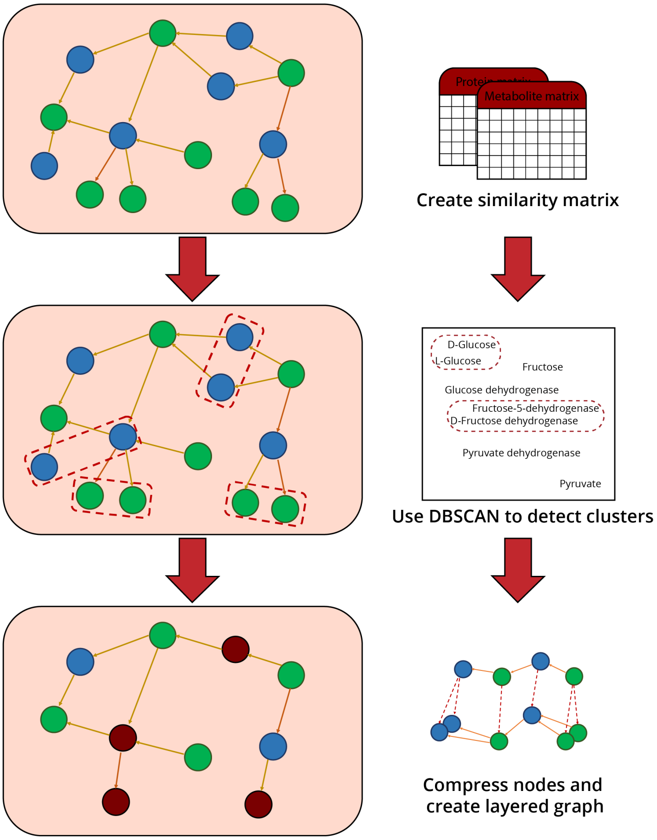


Figure 6.5: Graph compression prevents redundancy. All protein and metabolite names are subjected to similarity comparison, and similarity matrices, are created for both molecule types. Clustering is then applied to similarity matrices and clusters are used to construct a compressed graph that reduces network complexity.

rus species was obtained with keywords ”biotic stress response”, ”*Citrus sinensis*”, ”*Citrus clementina*”, ”*Arabidopsis thaliana*” and ”*Physcomitrella patens*”. The two last species are model organisms for plants and stress responses, respectively. The novel pathways were evaluated by Gene Ontology enrichment analysis of recovered genes and by manual comparison to the literature. Additionally, the citrus pathway was evaluated as a stress response gene-set in the enrichment analysis of the transcriptional response of citrus to antibiotic treatment [57].

Table 6.5: EcoCyc pathways used for the analysis of E. coli pathway reconstruction with the total number of relationships present in each pathway.

Pathway	Number of relationships
Pantothenate and CoenzymeA	22
Arginine and polyamine biosynthesis	50
Aspartate Superpathway	73
Superpathway of Chorismate metabolism	159
Enterobactin biosynthesis	26
Galactose degradation (Leloir pathway)	16
Superpathway of Ornithine degradation	22
Pentose Phosphate Pathway	20
Peptidoglycan Biosynthesis I	29
Polymyxin resistance	24
Superpathway of L-serine and glycine biosynthesis I	12
Superpathway of tetrahydrofolate biosynthesis	26
UDP-N-acetylmuramoyl-pentapeptide biosynthesis I	21

6.2.10 Padhoc installation and utilization

Padhoc is publicly available at https://github.com/ConesaLab/padhoc and runs in Linux systems with Python v2.7. Padhoc requires the prior installation of TEES, Metrecon, Neo4j and tmChem. Guidelines for the installation of these dependencies can be found at Padhoc's download site. Padhoc is used by running the run_padhoc.py script with a list of keywords that represent the pathway to search for and the organism of interest. After Padhoc finishes extracting the text and feeding the Neo4j database, the graph is compressed using the script compress_graph.py. After these two steps are completed, the network will be available at the user's local Neo4j database. More details regarding how to run Padhoc, as well as the examples used in this Chapter, can be found at Padhoc's download site.

6.3 Results

6.3.1 Building a user-defined pathway with Padhoc

Figure 6.1 shows Padhoc's scheme for creating a user-defined pathway. Essentially, the procedure starts with a set of keywords that define the pathway the user wishes to create, together with an organism name. The organism name is used to query a compendium of databases (Table 6.1) to retrieve all available metabolic and signaling information for that species, including entity names for genes, metabolites and proteins (and their synonyms), reactions, protein-protein interactions and transcription factor (TF)-protein interactions. Molecules are assigned a UniProt and CheBI ID, and all information is stored in a Neo4j graph database, where nodes represent molecular entities and edges represent the relationships between them (Figure 6.1, left).

The input set of keywords, together with the organism name, are used to query the scientific literature to retrieve PubMed IDs (PMIDs) and their associated text. Alternatively, a list of PMIDs or text can be supplied as input for Padhoc. TEES and Metrecon programs are then used, in combination with the NER engines BANNER and tmChem, to extract protein names, metabolites and reactions from text. Once the literature data has been extracted, entity names are assigned a UniProt/CheBI ID, when possible, and are then added to the Neo4j database. In this process, if a text-mining identified entity ID was already present in Neo4j, text-mining derived relationships are incorporated and associated to the existing IDs, otherwise new entity IDs and their relationships are added to the database (Figure 6.1, right). Entity names stored in Neo4j are compared to each other and clustered by semantic similarity to create 'compressed' nodes that collapse redundant information while maintaining links to source nodes. Finally, when multiple species are submitted to Padhoc (i.e., a non-model organism and a related model species), InParanoid is used to establish orthologous relationships and create hybrid pathways that gather relevant information from both species.

Once the combination of established knowledge (obtained from databases) and emerging knowledge (obtained from text-mining) are combined in the Neo4j platform, the targeted pathway is retrieved using Cypher queries on the Neo4j database. The resulting biological network can be filtered according to the number of appearances of each entity in the text or can be manually modified by the user. Additionally, customized queries can be used on the newly constructed pathway, for example, to recover a set of entities of interest (e.g., a list of genes or metabolites).

6.3.2 Information content of Padhoc pathways

Padhoc pathway data are stored in a Neo4j graph database and consist of four types of nodes and ten types of relationships. Nodes are either “Protein”, “Enzyme” or “Compound”, representing the primary structure of the graph (Figure 6.6A), or “Compressed”, which corresponds to a compacted version of semantically similar nodes (Figure 6.6B). Every node is assigned a stable ID, which generally corresponds to their UniProt/ChEBI identifier, although some stable IDs use the text name if the entity could not be matched to any database ID. The ten types of relationships include: “TM_relationship”, “Brenda_database”, “StringDB_interaction”, “TF_regulation”, “Pazar_relationship”, “IntAct_interaction”, “Omnipath_interaction”, “Orthology_relationship”, “compressed_to” and “compressed_relationship” (Table 6.3). The first eight relationship types define the primary information content of the pathway, while the last two are part of the condensed graph.

Pathway links contain properties that store information extracted from the text and databases. Brenda edges contain detailed information of the reaction, the Enzyme Code (EC) of the enzyme that drives the reaction and the species where this reaction was found. Text mining properties include the training dataset used by TEES (corpora), the type of reaction, the number or appearances in text (nbs), the search query used and the sentences where the reaction was extracted from the text. Omnipath relationships include post-translational modifications, while the Intact database includes the detection method used to identify interactions, the confidence score, the interaction method and type. In

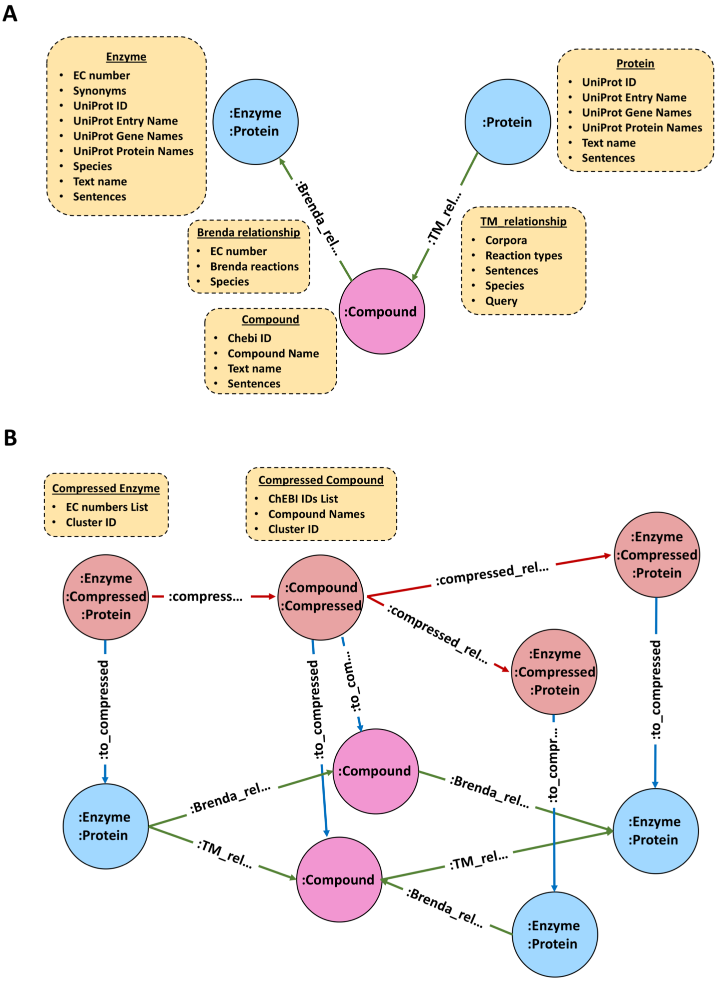


Figure 6.6: (A) Neo4j structure of connection between nodes by different types of relationships. Nodes are labelled according to their represented molecular type, while relationships are labelled based on their source information. Both nodes and edges have properties that are stored in Neo4j. (B) Neo4j structure of a compressed graph, with connections to the primary graph structure.

the compressed relationship, only sentences are included as a property. Properties stored in each node and relationship types are detailed in Tables 6.2 and 6.3.

6.3.3 Pathway validation using *E. coli* pathways

Since Padhoc was conceived to create novel pathways, a direct validation of the method is challenging. Therefore, we first evaluated whether Padhoc was able to faithfully recapitulate existing pathway data by comparing Padhoc results with curated pathways from established databases. A total of 13 *E. coli* pathways, representing a wide range of cellular processes of different complexity, were selected from the EcoCyc database [93] for evaluation. The scientific literature reported by EcoCyc as an information source for these pathways was used as input for our methodology. Table 6.5 lists the 13 biological pathways used for this analysis, as well as the number of relationships that comprise each pathway. Figures 6.7A and B show the sensitivity and specificity of the method as a function of the supporting evidence (number of appearances in text) for nodes and relationships.

For a support threshold of more than one sentences, Padhoc recovered around 65% of the reactions in the EcoCyc pathways, although only 15% of the reactions in the Padhoc pathways were also present in reference database. For two supporting sentences, mean sensitivity values were 51%, and specificity substantially improved to 24% (details for all pathways are provided in Table 6.6). These results suggest that, while Padhoc was able to recover most of the elements present in the reference E. coli pathways, a large number of additional entities and links were included when compared to the EcoCyc database. To determine whether new discovered relationships in these pathways were missing pathway information or false additions, Padhoc relationships absent from EcoCyc were manually evaluated and curated for the Pantothenate and Coenzyme A biosynthesis pathway. Relationships absent from the EcoCyc pathway were given a score (rank) representing the quality of their relevance to the pathway. Rank 1 was used for relationships that were recovered from text but either were extacted incorrectly by Padhoc or represented descriptions that were not relevant to the pathway (e.g., RNA polymerase reactions when describing molecular biology methods). Relationships were assigned rank 2 when text was recovered correctly, but the relationship to the pathway was unclear (e.g., folK gene,

Table 6.6: Analysis of pathway reconstruction using Padhoc, applying filters from 1 to >5 supporting mentions in text .

		$>$**0**			$>$**1**			$>$**2**	
	Rels	**Rels_DB**	**% similarity**	**Rels**	**Rels_DB**	**% similarity**	**Rels**	**Rels_DB**	**% similarity**
Pantothenate and CoenzymeA	228	22	81,81%	88	22	81,81%	44	22	72,72%
Arginine and polyamine biosynthesis	609	50	0.64	229	50	0.42	67	50	0.18
Aspartate Superpathway	1133	73	60,27%	530	73	58,90%	234	73	46,57%
Superpathway of Chorismate metabolism	1268	159	53,45%	609	159	52,20%	351	159	47,79%
Enterobactin biosynthesis	269	26	38,46%	105	26	38,46%	57	26	34,61%
Galactose degradation (Leloir pathway)	290	16	81,25%	94	16	81,25%	46	16	0.75
Superpathway of Ornithine degradation	169	22	27,27%	58	22	13,63%	28	22	13,63%
Pentose Phosphate Pathway	662	20	0.8	289	20	0.8	138	20	0.7
Peptidoglycan Biosynthesis I	294	29	62,06%	99	29	62,06%	53	29	62,06%
Polymyxin resistance	141	24	37,50%	31	24	29,16%	6	24	0.25
Superpathway of L-serine and glycine biosynthesis I	229	12	0.75	79	12	0.75	34	12	0.5
Superpathway of tetrahydrofolate biosynthesis	497	26	0.73	182	26	0.65	60	26	0.5
UDP-N-acetylmuramoyl-pentapeptide biosynthesis I	267	21	76,19%	77	21	76,19%	42	21	76,19%
		$>$**3**			$>$**4**			$>$**5**	
	Rels	**Rels_DB**	**% similarity**	**Rels**	**Rels_DB**	**% similarity**	**Rels**	**Rels_DB**	**% similarity**
Pantothenate and CoenzymeA	14	22	0.5	7	22	36,36%	1	22	0.09
Arginine and polyamine biosynthesis	30	50	0.1	8	50	0.04	7	50	0.04
Aspartate Superpathway	120	73	45,20%	53	73	26,02%	23	73	12,32%
Superpathway of Chorismate metabolism	224	159	37,10%	112	159	22,64%	90	159	16,98%
Enterobactin biosynthesis	29	26	34,61%	23	26	30,76%	13	26	19,23%
Galactose degradation (Leloir pathway)	25	16	62,50%	21	16	62,50%	18	16	62,50%
Superpathway of Ornithine degradation	3	22	4,50%	0	22	0	0	22	0
Pentose Phosphate Pathway	75	20	0.7	49	20	0.7	41	20	0.7
Peptidoglycan Biosynthesis I	24	29	58,62%	21	29	58,62%	17	29	51,72%
Polymyxin resistance	3	24	0.25	3	24	0.25	3	24	0.25
Superpathway of L-serine and glycine biosynthesis I	9	12	33,33%	5	12	0.25	2	12	0
Superpathway of tetrahydrofolate biosynthesis	27	26	0.5	18	26	26,92%	12	26	23,07%
UDP-N-acetylmuramoyl-pentapeptide biosynthesis I	27	21	76,19%	20	21	76,19%	18	21	66,66%

which interacts with 2-amino-6-hydroxymethyl-7,8-dihydropteridine-4-ol as part of tetrahydrofolate biosynthesis pathway). Finally, rank 3 was assigned to relationships that extended the EcoCyc pathway in a meaningful way.

Figure 6.7C shows that over 40% of the added relationships were assigned rank 3, indicating that a significant number of novel relationships added relevant knowledge to the pathway. Moreover, rank 3 sentences showed a slightly higher number of appearances in text than relationships with inconclusive relevance for the pathway (Figure 6.7D), while sentences with no relevance were barely supported. To identify a suitable support filter, we calculated the rate of bona fide pathway components -i.e., the sum of the rank 3 novel discoveries and the recovered pathway elements when compared with EcoCyc- as a function of the number of occurrences in the text (Figure 6.7E). Rank 1 novel discoveries were considered confirmed false positives. As expected, the rate of correctly assigned pathway elements increased with the support, and with just more than two supporting sentences this rate reached 79% of the pathway elements, while 14% were inconclusive (rank 2) and 7% are confirmed false calls (Figure 6.7E). Interestingly, >3 supporting sentences resulted in 100% correct assignment of pathway features.

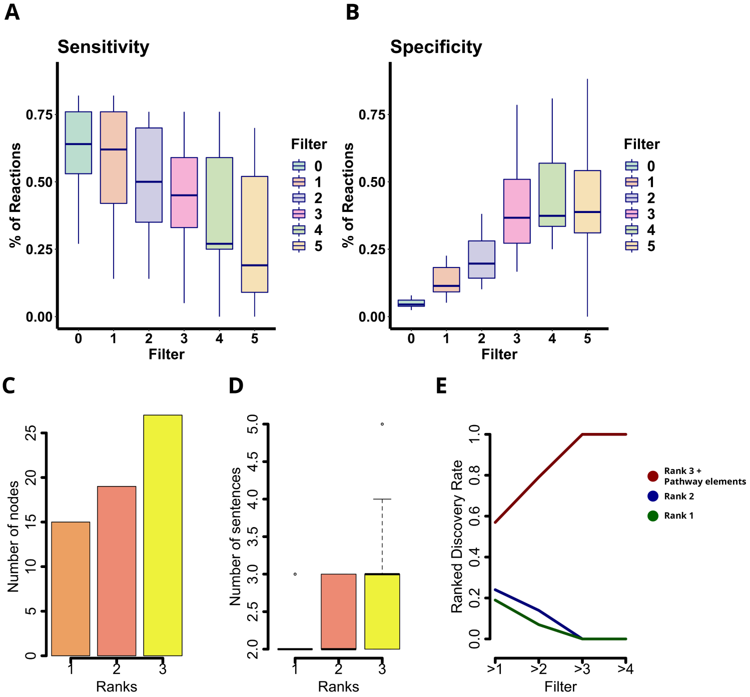


Figure 6.7: Padhoc evaluation with E. coli pathways. Specificity (A) and sensitivity (B) of reconstructed relationships from E. coli pathways compared to EcoCyc curated database. (C-E) Manual curation of the Pantothenate and CoenzymeA biosynthesis pathway. (C) Number of relationships assigned to each quality rank. (D) Number of sentences supporting each rank assignment. (E) Percentage of confirmed, inconclusive and false pathway calls as a function of the number of supporting sentences

6.3.4 Reconstruction of Human Histone Acetylation Pathway

The ability of Padhoc to reconstruct novel pathways was assessed with the reconstruction of the human histone acetylation pathway, in which metabolic generation of acetyl-coA supplies the transference of acetyl groups to histone tails through the activity of histone acetyltransferases (HATs). This pathway is gaining scientific interest as it represents an important component of the metabolic control of the epigenetic changes that regulate gene expression, as has been sown in previous chapters. Although several reviews have already been published [21, 140, 198], the information is yet to be included in pathway databases. We used Homo sapiens and ”histone acetylation” as keywords for Padhoc and obtained 38 research articles that were used for pathway reconstruction.

Padhoc obtained a network comprised of 562 features (183 genes, 26 metabolites and 353 connections) that were supported by more than one text entry (Table 6.7). GO enrichment analysis of the recovered genes returned numerous terms related to histone binding, fatty-acid metabolic process, epigenetic regulation of gene expression, regulation of gene silencing and acetyl-coA biosynthetic processes, among others, suggesting a successful recovery of gene functions relevant to the targeted pathway. For illustration purposes, we show in Figure 6.8 the pathway representation after applying a support threshold of 3 or more sentences, which retains the most relevant elements of this pathway.

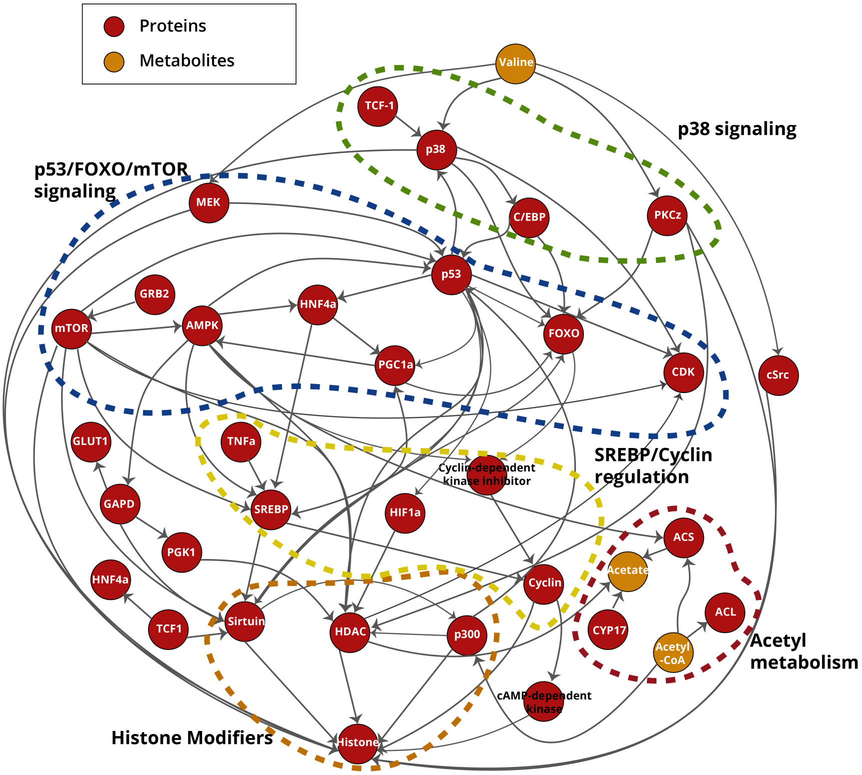


Figure 6.8: Homo sapiens Histone Acetylation pathway. Five main subpathways are represented in this network: two signaling pathways (p53, mTOR and FOXO and p38); one transcriptional regulation (SREBP); acetyl-CoA metabolic pathways; and histone modification mechanisms.

In agreement with the literature, the pathway recapitulates the connection of acetyl-CoA and acetate with Histone Acetyltransferases HATs (in this case

Table 6.7: Total number of Proteins and Compounds in the Homo sapiens histone acetylation pathway, using filters from 1 mention to >3 mentions in text.

Filter	**Protein**	**Compound**	**Total**
No filter	777	92	869
$>$**1**	183	26	209
$>$**2**	66	9	75
$>$**3**	28	4	32

p300) and histone deacetylases (HDACs, represented by HDAC and Sirtuin), to supply the acetyl group required to acetylate/deacetylate histones, respectively (circled in orange) [21, 197]. At central positions of the pathway, a highly connected signaling network is observed (circled in blue) including, among other kinases, CDK and mTOR ,which are part of the phosphorylation cascade that activates HDACs to regulate their function as histone modifiers [25, 124], and transcription factors p53 and FOXO, which regulate transcription of growth and apoptotic genes [204]; p53 is, in turn, acetylated by p300 [157] to regulate its activity. The pathway also shows the kinase activation of other transcription factors controlling the expression of genes involved in nutrient response, such as C/EBP (circled in green), which is regulated by p38 [194] and SREBP (circled in yellow) and controlled by mTOR activity [164]. These factors modulate acetyl-CoA availability by regulating the expression of several metabolic genes (circled in red) and are therefore indirect regulators of the acetylation of histones [165, 170]. In summary, Padhoc recovered an integrated pathway that joins the metabolic component of HAT and HDAC-mediated histone acetylation with the signaling cascade that activates both histone acetylases and the metabolic enzymes linked to acetyl-CoA metabolism.

6.3.5 Padhoc for pathway reconstuction in non-model species

One of the distinct signatures of Padhoc's framework is the incorporation of the InParanoid homology search functionality to facilitate pathway construction for non-model organisms. We evaluated this functionality by using Padhoc to obtain the plant biotic stress response pathway for Citrus sinensis. Citrus are well-studied plants, with available genomic sequences [202], for which relevant literature on biotic and abiotic stress is available [122]. However, to our surprise, the molecular map of these stress response mechanisms in citrus -as in other plants- is poorly represented in public databases. In KEGG, the closest pathway is the MAPK signaling pathway, which describes plant stress sensing; moreover, in PlantCyc [208], there are a total of 498 pathways for *Citrus sinensis*, but none of them are specifically associated to “stress response”. Similarly, the stress response for the model plant species *Arabidopsis thaliana* is poorly represented in public databases with the same type of limitations.

Feeding Padhoc with the keywords Citrus sinensis, Citrus clementina, Arabidopsis thaliana, Physcomitrella patens (a fungal model organism for stress responses) and ”biotic stress response”, a total of 150 papers were recovered, from which Padhoc extracted 69 genes, 43 metabolites and 111 reactions with more than one supporting sentence. Genes were enriched in GO terms related to oxidative functions, hormone signaling, response to different stresses, catalytic activities and cellular fluxes. Figure 9 shows the citrus stress response pathway obtained by Padhoc with a filter of two or more supporting sentences. The Neo4j network representation reveals a comprehensive picture of the different molecular events that take place under stress conditions in plants. We observed a cluster of genes and metabolites representing glutathione metabolism, one of the most important detoxification pathways in plants and animals [69]. The network also includes numerous metabolic and cellular responses associated to stress. For example, PMD, ADPG genes and pectins are involved in cell-wall plasticity during adaptation to stress conditions [66]; Acetyl-CoA metabolism is known to be down regulated in response to stress [49]; ACP mediates lipid signaling and metabolism under biotic and abiotic stress [192]; SPS and SS genes are part of the control of sucrose and starch accumulation that follows stress [79, 114]; while many amino-acids accumulate upon stress conditions to act as osmolytes, regulate ion transport and modulate detoxification [148].

Also important in the Padhoc network is the connection of different plant hormone signaling pathways, including abscisic acid (ABA), jasmonic acid (JA), and salicylic acid (SA), which are part of plant defense signaling systems [177], described with different granularity levels both in both KEGG and PlantCyc. These hormone pathways are represented in the compressed and filtered Padhoc network by a few members of their signaling cascade. Double-clicking any of these compressed nodes recovers the underlying information available, thanks to the inclusion of previous pathway data in the Neo4j database (Figure 9, orange and blue nodes). This revealed new components of the ABA and SA biosynthesis from xanthosine and benzoate, respectively. Node de-compression also revealed that both sub-pathways are connected by methyltransferases that control enzymatic activity and share S-Adenosyl-Methionine (SAM) as substrate. Moreover, the salicylic acid pathway is further linked to the glutathione pathway by the utilization of acetate groups as substrate for SFGH esterase (Figure 6.9). These results demonstrate the power of Padhoc, not only for joining literature with established knowledge to construct tailored pathways, but also in connecting different branches of complex molecular circuits.

Finally, we evaluated this pathway in the context of a gene expression analysis. We used a recent study that evaluated the response of the orange tree to Benzbromarone, a new antibiotic proposed for the treatment of citrus greening [57]. Citrus greening is caused by the bacteria *Liberibacter asiaticus*, which affects the phloem of the tree, leading to dramatic reductions in yields and eventually plant death. In this study, the trunks of affected orange trees were injected with a Benzbromarone infusion to treat infection; leaves were collected after several weeks to evaluate the transcriptional response of the tree to the antibacterial treatment and thereby understand possible degradation mechanisms for the drug in the plant. The study found 404 genes to be deferentially expressed with respect to mock-treated controls. Enrichment analysis using several pathway resources indicated the activation of sucrose, lignin and cell-wall related pathways, but no clear insights were obtained regarding stress response mechanisms [57]. We used the Padhoc citrus stress response pathway as a gene-set for enrichment analysis of the differentially expressed genes. We found significant enrichment for pathway representations at $>$1 support (p-value = 0.014), indicating that a significant stress response was transcriptionally active.

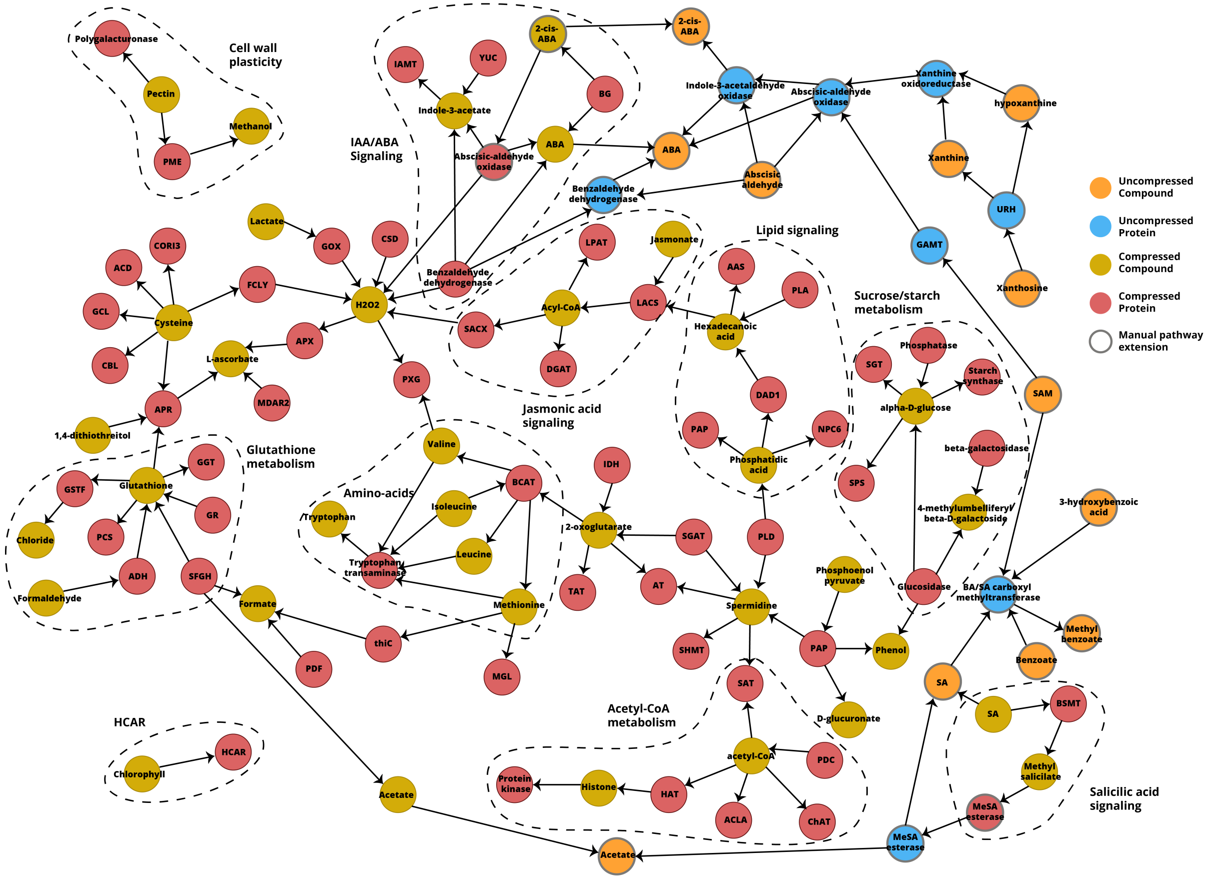


Figure 6.9: Biotic stress response pathway in Citrus sinensis obtained by Padhoc. Pathway elements have more than one text supporting sentence.

6.4 Discussion

Biological pathway and molecular interaction databases have evolved as information technology resources to become critical tools in supporting systems biology research, where the assessment of molecular relationships is a fundamental component of the knowledge discovery process. However, the growth of pathway databases is not only sustained by the advance of the computational sciences, but also thanks to the incorporation of a labor- and time-consuming manual curation process that guarantees the quality and completeness of the pathway models. Although this expert review is important to ensure the utility of pathway resources, it has the downside of resulting in design and coverage restrictions that may limit the application of pathway-based analysis methods to emerging biological domains. Padhoc development was motivated by the realization of the limitations of existing databases for providing adequate data to support pathway analysis in a number of recent studies from our lab.

Padhoc combines knowledge from established pathway databases with a text mining approach to retrieve the molecular components and interactions for any pathway of interest. This allows for the recovery of most state-of-the-art data for biological processes that are under active research, while guaranteeing a curated skeleton of molecular interactions at the base of pathway reconstruction. Padhoc stores the molecular relationships in Neo4j [130], a graph database that offers a user-friendly interface and flexibility to manipulate the pathway. Padhoc's performance was tested on the reconstruction of a number of Escherichia coli pathways from the EcoCyc database. The assessment of the reconstruction reveals that a large fraction of the recovered pathway features faithfully recapitulates information from the available literature, but also that Padhoc networks did not capture all the current knowledge on the targeted pathway. This was evident in reconstructing *E.coli's* Pantothenate and Co-A biosynthesis pathway, where verified calls were nearly 80% of the inferred elements, but 19% of the EcoCyc pathway components were missing. Since both BANNER and tmChem have shown to provide $>$90% success in entity recognition from text [109, 110], this result may indicate that some interactions included in early reference pathways cannot be traced back from the electronically available manuscripts. This limitation should not be as critical for new research domains and is expected to decrease with time as publication of fully open access manuscripts, amenable to text mining, becomes a general practice. The manual evaluation of the *E. coli* Pantothenate and Co-A biosynthesis pathway also revealed that an important amount of bona fide new pathway elements were added by the text mining algorithm. This means that, even for well-established pathway maps, Padhoc was able to provide meaningful updates. At this point of Padhoc development, we have not included the reference pathway data in the reconstruction pipeline, as this would undermine our ability to evaluate the expected performance of Padhoc for newly proposed pathways. However, the flexible structure of the software, where multiple databases are combined and integrated into a unified feature ID system, would make such extensions feasible. Padhoc could also benefit from new NER mechanisms, such as HUNER [195], and from pathway export into SBML or BioPAX formats, which will be considered in future versions of the software.

To evaluate Padhoc's capability for constructing pathways not yet available in databases, we used the pipeline to create the histone acetylation pathway in *Homo sapiens* and the biotic stress response in *Citrus sinensis*, which also represent analysis scenarios for model and non-model organisms, respectively. In both cases, results showed that Padhoc is able to connect multiple processes that contribute to establish biological functionalities to be modelled as a pathway. Histone acetylation is driven by metabolic processes that lead to acetyl-CoA accumulation and by signaling cascades that control the enzymatic activity of histone modifiers [21, 197]. Although molecular connections that drive histone acetylation can be found in pathway databases, there is no pathway that combines the different mechanisms involved. Padhoc was able to successfully integrate at least five different functional components that contribute to the modification of histones and provided a joint view of this cellular process. Also, in the case of the citrus stress pathway, Padhoc integrated the metabolic, detoxification and hormone signaling aspects, generating a comprehensive representation of this response. This is a unique property of our approach, as the definition and number of the keywords used to run Padhoc provide a great deal of flexibility to establish pathway boundaries as a function of the researcher's needs. Another element of flexibility and versatility of Padhoc is achieved thanks to the utilization of the graphical database Neo4j to host the pathway data. This facilitates, for example, adjustment of the support threshold to control the confidence and extension of the pathway map. Similarly, the ability to compress/uncompress pathway nodes allows for high resolution and navigation on particular aspects of the retrieved network, while uncovering hidden links among its components. This interactive process is particularly useful when using pathway insights for hypothesis generation and mechanistic interpretation of the data.

Finally, a distinctive characteristic of Padhoc is the support it provides for pathway reconstruction in non-model organisms, thanks to the integration of the InParanoid database and homology search functions. Development of pathway maps for non-model species is usually delayed with respect to model organisms, which imposes a disadvantage to researchers working in these fields. Therefore, the availability of an easy-to-use tool for on demand pathway construction in these species is particularly useful. We showed that Padhoc successfully created a relevant stress pathway in *Citrus* that was effective in providing a suitable gene-set for the analysis of the transcriptional response of the orange tree to antibiotic treatment against citrus greening disease.

Chapter 7

Conclusions

In this thesis, we used multivariate statistical methods for the integrative analysis of different types of high-throughput molecular data to study the impact of metabolic changes on gene expression regulation. Gene expression is a tightly regulated process with multiple layers of molecular control that involve chromatin state (histone modifications and DNA methylation), active transcriptional regulation by transcription factors and other components of the transcriptional machinery, and post-transcriptional regulation, where splicing, transport and microRNAs modify the steady-state levels of mature RNAs. Metabolite levels are also known to be involved in the control of gene expression and to have an important role in the control of biological rhythms, as metabolic oscillations coordinate with gene expression oscillations to define biological clocks. Although it is generally accepted that the metabolic control of gene expression is a chromatin modification-meditated process, the mechanisms by which metabolic reactions ultimately translate into control of the epigenetic landscape are still unknown. Here, we used the Yeast Metabolic Cycle (YMC) as a model system with which to study the impact of metabolism on chromatin changes and to investigate the interplay between epigenetics and transcription factor activity to control gene expression regulation.

Firstly, we used existing YMC RNA-seq and histone modification ChIP-seq datasets to determine which histone modifications correlate best with gene expression, and we functionally characterized the processes where they are involved. Secondly, we obtained and processed an available NET-seq dataset that covered the totality of the cycle. Thirdly, we generated new ATAC-seq and metabolomics datasets that matched previous RNA-seq, NET-seq and ChIP-seq datasets to create a multi-omics dataset that spanned transcriptional, epigenetic and metabolic layers of this system with high resolution across the cycle. Then, we applied multivariate statistics to this multi-omics dataset to study the relationships among these three molecular layers to propose models of interaction and cross-regulation.

Finally, we addressed an important problem for the interpretation of multi-omics data integration results, which is the availability of pathway models that incorporate the most state-of-the-art information to serve as templates for enrichment and functional analyses. We developed Padhoc, a bioinformatics tool that creates ‘ad hoc’ biological pathways, tailored to user needs. Padhoc combines curated information from pathway databases and new molecular data extracted from scientific articles. In Chapter 6, we demonstrate the utility of this tool for inferring molecular networks that represent the biological processes involved in the connection between metabolism and epigenetic modifications and for creation of pathways for non-model organisms.

The conclusions of this thesis are stated below, and they summarize the goals set in our Chapter 2:

1) **Understand the impact of chromatin in gene expression across the Yeast Metabolic Cycle.**

- We obtained RNA-seq and ChIP-seq datasets for 8 histone marks from the public domain and integrated them to create a multi-omics dataset with matching time-points throughout the Yeast Metabolic Cycle. Gene expression data was fit into three clusters that matched the three YMC functional phases and recapitulated results from previous studies, which validated our preprocessing pipeline.
- Analysis of the relationship between the histone modifications and gene expression using N-PLS revealed that H3K9ac and H3K18ac histone marks had the highest impact on gene expression changes. This analysis also revealed that both chromatin regions had a similar impact on gene expression variability, and that the impact of histone marks is associated with gene variability in two phases.
- Using MORE GLMs we determined the histone modifications that were more closely related to the transcriptional changes of each individual gene and the biological processes that each histone mark is more associated with. H3K9ac and H3K18ac were linked to fatty acid biosynthesis in RB, all histone modifications were associated to Pentose phosphate pathway in OX and to TCA in RB and RC phases; and H3K14ac, H3K56ac, H3K9ac and H3K18ac were closely associated with fatty acid oxidation in RC phase.
- MORE GLMs also allowed us to decipher the interplay between YMC functional phases and histone mark control, revealing that H3K9ac displayed higher control at OX phase, RB was linked to H3K56ac control and RC was associated with the H3K18ac mark.
- Using a bioinformatics approach, we detected 13 transcription factors that are likely to be relevant for the regulation of the cycle and proposed a pro-

gram of specific TF-histone mark interactions that regulate the transcriptional oscillations in the YMC.

- Association analysis between histone modifications and TFs suggested that Pip2 and Hfi1 cooperate with H3K9ac and H3K18ac to drive the OX phase.

2) Extract metabolomics and ATAC-seq data to obtain a multi-omics dataset that match previous RNA-seq and ChIP-seq data in the YMC.

- We designed a protocol to accurately extract metabolomics and ATAC-seq datasets from the YMC, that match previous datasets available for this system.
- We obtained 21 metabolomics samples in quintuplicate, which covered the three YMC phases, and measured a total of 60 metabolites in each sample. 18 samples were processed from the fermenter in two batches to measure genome-wide chromatin accessibility (ATAC-seq).
- ATAC-seq preprocessing suggested that yeast cells undergo physiological changes during the YMC that make the chromatin of the cells less accessible to transposition reactions. Extra efforts should be made in the future to obtain higher resolution in late RB phase of the cycle to obtain a better characterization of the chromatin dynamics.
- Clustering of the differential abundant features from the datasets revealed that three clusters better captured the metabolomics variability, while the two clusters better described the oscillations in chromatin accessibility. These results suggested that metabolic oscillations match gene expression dynamics. Chromatin dynamics offer a different behavior that resembles oxygen oscillations, suggesting that chromatin acts as a sensor the external conditions.
- Chromatin accessibility was used for transcription factor discovery, highlighting the activity of Azf1, Sum1, Abf1 or Rsc3. These TFs are the main candidates for coordination of gene expression regulation with epigenetic changes.

3) Accurately model how metabolic signaling affects gene expression through its impact in the YMC epigenetics landscape.

- The extraction of ATAC-seq and metabolomics data and the recovery of existing RNA-seq, histone modification and NET-seq datasets, allowed us to prepare an extensive multi-omics dataset for the YMC.
- Modelling gene expression changes as a response variable of histone modifications and ATAC-seq revealed a two-phase gene expression response to epigenetic variability, where RB phase shows the lowest variation. This matches the profiles of the two epigenetic datasets and confirms a two-phase epigenetic regulation of the YMC.
- Using metabolites as an explanatory variable for gene expression oscillations in PLS revealed that there are two groups of metabolites in each YMC phase, which divided into metabolic clusters of early and late accumulation within their respective phase. The functional analysis of these metabolites highlighted groups of metabolites, such as NAM-derivatives, that have a coordinated function in epigenetic changes.
- The application of PLS-PM allowed us to create a multivariate model using the multi-omics dataset. We obtained a testable configuration of the model that allowed the selection of features that have an impact in the cycle.
- The PLS-PM model confirmed the presence of two stages of metabolic accumulation, where early-accumulated metabolites are better explained by the gene expression oscillations, and late-accumulated metabolites have a more direct impact in epigenetic regulation.
- Using information from the PLS-PM model, we detected several processes that are known to connect metabolism with epigenetic regulation:
  - ATP accumulated in OX phase, where Ino80 ATP-dependent chromatin remodeler activity has been previously reported. This suggests that ATP accumulation (high-energy state) triggers the remodeler activity to activate amino acid biosynthesis.

  - Acetyl-CoA was linked to OX phase acetylation, as has been reported in previous studies, targeting growth genes through the H3K9ac histone mark.
  - NAD accumulation in early RC phase led to sirtuin regulation, explaining accumulation of H3K18ac marks and possibly regulating fatty acid degradation.
  - There was an accumulation of metabolites that have an impact in methylation (methionine, alpha-ketoglutarate, etc.), thus it would be of interest to include a greater range of histone marks in the analysis to aid the functional interpretation of the metabolic activity.
- PLS-PM model linked Aspartate and Phenylalanine with RC-phase regulation. Although these metabolites have no direct impact in epigenetic mechanisms, it would be interesting to further study their role in the YMC progression.

4) **Develop a bioinformatics framework in which to create up-to-date pathways tailored to user needs.**

- We developed Padhoc, a framework that combines pathway database information with recent findings extracted from scientific articles. This combined information is successfully stored and visualized in the Neo4j graph database, which also aids in the task of manual curation through the storage of properties for entities and relationships.
- Padhoc's compression algorithm successfully clusters similar nodes and creates a condensed graph, which simplifies the pathway while retaining all relevant information.
- We included an orthology module that allows for quering of multiple species in Padhoc and connects orthologous proteins between the different species. This orthology search facilitated reconstruction of pathways from non-model organisms using the information from the closest model species.

- Padhoc was tested using 13 existing *E. coli* pathways. Sensitivity/Specificity analyses and posterior manual curation revealed that Padhoc can extract most information present in the original pathway and complete it with relevant relationships that integrate all knowledge around the pathway in question.
- Successful reconstruction of the *H. sapiens* histone acetylation pathway suggests that Padhoc can be used for the creation of signaling pathways that are underrepresented in pathway databases.
- Padhoc was also used to create a stress-response pathway for *C. sinensis*, a non-model species. Padhoc's usage in obtaining such complex networks could offer support for functional interpretation of omics datasets in organisms that currently lack trustworthy resources.

Appendix 1:
List of indicators from PLS-PM Latent Variables

Table 7.1: Top 20 indicators from Ox Gene Expression LV, with description.

Gene ID	Description
YJL192C	ER-membrane protein; subunit of evolutionarily conserved EMC (Endoplasmic Reticulum Membrane Complex) implicated in ERAD (ER-associated degradation) and proper assembly of multi-pass transmembrane (TM) proteins; EMC acts in yeast as an ER-mitochondria tether that interacts with outer membrane protein Tom5 of TOM (Translocase of the Mitochondrial Outer Membrane) complex; suppressor of pma1-7, deletion of SOP4 slows down export of wild-type Pma1p and Pma1-7 from the ER
YGR040W	Mitogen-activated protein kinase (MAPK); involved in signal transduction pathways that control filamentous growth and pheromone response; regulates septum assembly, and may directly phosphorylate Bni4p; the KSS1 gene is nonfunctional in S288C strains and functional in W303 strains
YER118C	Transmembrane osmosensor for filamentous growth and HOG pathways; involved in activation of the Cdc42p- and MAP kinase-dependent filamentous growth pathway and the high-osmolarity glycerol (HOG) response pathway; phosphorylated by Hog1p; interacts with Pbs2p, Msb2p, Hkr1p, and Ste11p
YNL166C	Linker protein responsible for recruitment of myosin to the bud neck; interacts with the C-terminal extensions of septins Cdc11p and Shs1p and binds Myo1p to promote cytokinesis
YDR062W	Component of serine palmitoyltransferase; responsible along with Lcb1p for the first committed step in sphingolipid synthesis, which is the condensation of serine with palmitoyl-CoA to form 3-ketosphinganine
YJR143C	Protein O-mannosyltransferase; transfers mannose residues from dolichyl phosphate-D-mannose to protein serine/threonine residues; appears to form homodimers in vivo and does not complex with other Pmt proteins; target for new antifungals
YDR331W	ER membrane glycoprotein subunit of the GPI transamidase complex; adds glycosylphosphatidylinositol (GPI) anchors to newly synthesized proteins; human PIG-K protein is a functional homolog
YNL231C	Phosphatidylinositol transfer protein (PITP); controlled by the multiple drug resistance regulator Pdr1p; localizes to lipid particles and microsomes; controls levels of various lipids, may regulate lipid synthesis; homologous to Pdr17p; protein abundance increases in response to DNA replication stress
YPL050C	Subunit of Golgi mannosyltransferase complex; this complex mediates elongation of the polysaccharide mannan backbone; forms a separate complex with Van1p that is also involved in backbone elongation; this complex also contains Anp1p, Mnn10p, Mnn11p, and Hoc1p
YNL233W	Targeting subunit for Glc7p protein phosphatase; localized to the bud neck, required for localization of chitin synthase III to the bud neck via interaction with the chitin synthase III regulatory subunit Skt5p; phosphorylation by Slt2p and Kss1p involved in regulating Bni4p in septum assembly
YML061C	DNA helicase, potent G-quadruplex DNA binder/unwinder; possesses strand annealing activity; promotes DNA synthesis during break-induced replication; important for crossover recombination; translation from different start sites produces mitochondrial and nuclear forms; nuclear form is a catalytic inhibitor of telomerase; mitochondrial form involved in DNA repair and recombination; mutations affect Zn, Fe homeostasis; regulated by Rad53p-dependent phosphorylation in rho0 cells
YOR368W	Checkpoint protein; involved in the activation of the DNA damage and meiotic pachytene checkpoints; with Mec3p and Ddc1p, forms a clamp that is loaded onto partial duplex DNA; homolog of human and S. pombe Rad1 and U. maydis Rec1 proteins
YOL056W	Homolog of Gpm1p phosphoglycerate mutase; converts 3-phosphoglycerate to 2-phosphoglycerate in glycolysis; may be non-functional; GPM3 has a paralog, GPM2, that arose from the whole genome duplication
YML021C	Uracil-DNA glycosylase; required for repair of uracil in DNA formed by spontaneous cytosine deamination; efficiently excises uracil from single-stranded DNA in vivo; not required for strand-specific mismatch repair; cell-cycle regulated, expressed in late G1; localizes to mitochondria and nucleus
YDR295C	Subunit of the HDA1 histone deacetylase complex; possibly tetrameric trichostatin A-sensitive class II histone deacetylase complex contains Hda1p homodimer and an Hda2p-Hda3p heterodimer; involved in telomere maintenance; relocalizes to the cytosol in response to hypoxia
YDR235W	U1 snRNP protein involved in splicing; required for U1 snRNP biogenesis; contains multiple tetriatricopeptide repeats
YHR167W	Subunit of the THO and TREX complexes; THO connects transcription elongation and mitotic recombination, and TREX is recruited to activated genes and couples transcription to mRNA export; involved in telomere maintenance
YLR372W	Elongase; involved in fatty acid and sphingolipid biosynthesis; synthesizes very long chain 20-26-carbon fatty acids from C18-CoA primers; involved in regulation of sphingolipid biosynthesis; lethality of the elo2 elo3 double null mutation is functionally complemented by human ELOVL1 and weakly complemented by human ELOVL3 or ELOV7
YAR014C	Protein involved in bud-site selection; Bud14p-Glc7p complex is a cortical regulator of dynein; forms a complex with Kel1p and Kel2p that regulates Bnr1p (formin) to affect actin cable assembly, cytokinesis, and polarized growth; diploid mutants display a random budding pattern instead of the wild-type bipolar pattern; relative distribution to the nucleus increases upon DNA replication stress
YIL044C	ADP-ribosylation factor (ARF) GTPase activating protein (GAP) effector; involved in Trans-Golgi-Network (TGN) transport; contains C2C2H2 cysteine/histidine motif

Table 7.2: Top 20 indicators from Rb Gene Expression LV, with description.

Gene ID	Description
YLR390W	Protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
YPL078C	Subunit b of the stator stalk of mitochondrial F1F0 ATP synthase; ATP synthase is a large, evolutionarily conserved enzyme complex required for ATP synthesis; contributes to the oligomerization of the complex, which in turn determines the shape of inner membrane cristae; phosphorylated
YBL080C	Subunit of the trimeric GatFAB AmidoTransferase(AdT) complex; involved in the formation of Q-tRNAQ; mutation is functionally complemented by the bacterial GatB ortholog
YDL004W	Delta subunit of the central stalk of mitochondrial F1F0 ATP synthase; F1F0 ATP synthase is a large, evolutionarily conserved enzyme complex required for ATP synthesis; F1 translationally regulates ATP6 and ATP8 expression to achieve a balanced output of ATP synthase genes encoded in nucleus and mitochondria; phosphorylated
YNL026W	Component of the Sorting and Assembly Machinery (SAM) complex; the SAM (or TOB) complex is located in the mitochondrial outer membrane; the complex binds precursors of beta-barrel proteins and facilitates their outer membrane insertion; homologous to bacterial Omp85
YDR114C	Putative protein of unknown function; deletion mutant exhibits poor growth at elevated pH and calcium
YBR003W	Hexaprenyl pyrophosphate synthetase; catalyzes the first step in ubiquinone (coenzyme Q) biosynthesis
YPL059W	Glutathione-dependent oxidoreductase; mitochondrial matrix protein involved at an early step in the biogenesis of iron-sulfur centers along with Bol1p; hydroperoxide and superoxide-radical responsive; monothiol glutaredoxin subfamily member along with Grx3p and Grx4p
YML085C	Alpha-tubulin; associates with beta-tubulin (Tub2p) to form tubulin dimer, which polymerizes to form microtubules; relative distribution to nuclear foci increases upon DNA replication stress; TUB1 has a paralog, TUB3, that arose from the whole genome duplication
YJR133W	Xanthine-guanine phosphoribosyl transferase; required for xanthine utilization and for optimal utilization of guanine
YDR462W	Mitochondrial ribosomal protein of the large subunit; protein abundance increases in response to DNA replication stress
YMR267W	Mitochondrial inorganic pyrophosphatase; required for mitochondrial function and possibly involved in energy generation from inorganic pyrophosphate; human ortholog, PPA2, functionally complements the null mutant; mutations in human PPA2 cause a mitochondrial disease resulting in sudden unexpected cardiac arrest in infants
YKR065C	Constituent of the TIM23 complex; proposed alternatively to be a component of the import motor (PAM complex) or to interact with and modulate the core TIM23 (Translocase of the Inner mitochondrial Membrane) complex; protein abundance increases in response to DNA replication stress
YBL099W	Alpha subunit of the F1 sector of mitochondrial F1F0 ATP synthase; which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; F1 translationally regulates ATP6 and ATP8 expression to achieve a balanced output of ATP synthase genes encoded in nucleus and mitochondria; phosphorylated; N-terminally propionylated in vivo
YML081C-A	Subunit of the mitochondrial F1F0 ATP synthase; F1F0 ATP synthase is a large, evolutionarily conserved enzyme complex required for ATP synthesis; termed subunit I or subunit j; does not correspond to known ATP synthase subunits in other organisms
YCR023C	Vacuolar membrane protein of unknown function; member of the multidrug resistance family; YCR023C is not an essential gene
YPL127C	Histone H1, linker histone with roles in meiosis and sporulation; decreasing levels early in sporulation may promote meiosis, and increasing levels during sporulation facilitate compaction of spore chromatin; binds to promoters and within genes in mature spores; may be recruited by Ume6p to promoter regions, contributing to transcriptional repression outside of meiosis; suppresses DNA repair involving homologous recombination
YKL053C-A	Mitochondrial intermembrane space protein; forms complex with Ups2p that transfers phosphatidylserine from outer membrane to inner membrane for phosphatidylethanolamine synthesis; mutation affects mitochondrial distribution and morphology; contains twin cysteine-x9-cysteine motifs; protein abundance increases in response to DNA replication stress
YJR113C	Mitochondrial ribosomal protein of the small subunit; has similarity to E. coli S7 ribosomal protein
YOL012C	Histone variant H2AZ; exchanged for histone H2A in nucleosomes by the SWR1 complex; involved in transcriptional regulation through prevention of the spread of silent heterochromatin; Htz1p-containing nucleosomes facilitate RNA Pol II passage by affecting correct assembly and modification status of RNA Pol II elongation complexes and by favoring efficient nucleosome remodeling

Table 7.3: Top 20 indicators from Rc Gene Expression LV, with description.

Gene ID	Description
YDR003W	Vacuolar protein; presumably functions within the endosomal-vacuolar trafficking pathway, affecting events that determine whether plasma membrane proteins are degraded or routed to the plasma membrane; RCR2 has a paralog, RCR1, that arose from the whole genome duplication
YMR114C	Protein of unknown function; may interact with ribosomes, based on co-purification experiments; green fluorescent protein (GFP)-fusion protein localizes to the nucleus and cytoplasm; YMR114C is not an essential gene
YBL078C	Component of autophagosomes and Cvt vesicles; regulator of Atg1p, targets it to autophagosomes; binds the Atg1p-Atg13p complex, triggering its vacuolar degradation; unique ubiquitin-like protein whose conjugation target is lipid phosphatidylethanolamine (PE); Atg8p-PE is anchored to membranes, is involved in phagophore expansion, and may mediate membrane fusion during autophagosome formation; deconjugation of Atg8p-PE is required for efficient autophagosome biogenesis
YIL087C	Protein of unknown function; mitochondrial protein that physically interacts with Tim23p; null mutant displays reduced respiratory growth
YIL105C	Phosphoinositide PI4,5P(2) binding protein, forms a complex with Slm2p; acts downstream of Mss4p in a pathway regulating actin cytoskeleton organization in response to stress; TORC2 complex substrate and effector; protein abundance increases in response to DNA replication stress; SLM1 has a paralog, SLM2, that arose from the whole genome duplication
YPL247C	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and nucleus; similar to the petunia WD repeat protein an11; overexpression causes a cell cycle delay or arrest
YHR087W	Protein of unknown function involved in RNA metabolism; has structural similarity to SBDS, the human protein mutated in Shwachman-Diamond Syndrome (the yeast SBDS ortholog = SDO1); null mutation suppresses cdc13-1 temperature sensitivity; protein abundance increases in response to DNA replication stress
YEL060C	Vacuolar proteinase B (yscB) with H3 N-terminal endopeptidase activity; serine protease of the subtilisin family; involved in protein degradation in the vacuole and required for full protein degradation during sporulation; activity inhibited by Pbi2p; protein abundance increases in response to DNA replication stress; PRB1 has a paralog, YSP3, that arose from the whole genome duplication
YIL033C	Regulatory subunit of the cyclic AMP-dependent protein kinase (PKA); PKA is a component of a signaling pathway that controls a variety of cellular processes, including metabolism, cell cycle, stress response, stationary phase, and sporulation
YBR212W	RNA binding protein that negatively regulates growth rate; interacts with the 3’ UTR of the mitochondrial porin (POR1) mRNA and enhances its degradation; overexpression impairs mitochondrial function; interacts with Dhh1p to mediate POR1 mRNA decay; expressed in stationary phase
YBL086C	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cell periphery
YOR055W	Dubious open reading frame; unlikely to encode a functional protein, based on available experimental and comparative sequence data
YCR061W	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm in a punctate pattern; induced by treatment with 8-methoxypsoralen and UVA irradiation
YKL100C	Intramembrane aspartyl protease of the perinuclear ER membrane; acts in a branch of ER-associated degradation (ERAD) that degrades functional proteins rather than misfolded proteins; regulates abundance of high-affinity plasma membrane transporters, such as Ctr1p and Zrt1p, during the starvation response; has a presenilin fold; member of the GxGD family of intramembrane proteases; closest human homolog is signal peptide peptidase (SPP)
YJL144W	Cytoplasmic hydrophilin essential in desiccation-rehydration process; expression induced by osmotic stress, starvation and during stationary phase; protein abundance increases in response to DNA replication stress
YJL161W	Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
YML004C	Monomeric glyoxalase I; catalyzes the detoxification of methylglyoxal (a by-product of glycolysis) via condensation with glutathione to produce S-D-lactoylglutathione; expression regulated by methylglyoxal levels and osmotic stress
YCL038C	Vacuolar integral membrane protein required for efflux of amino acids; required for efflux of amino acids during autophagic body breakdown in the vacuole; null mutation causes a gradual loss of viability during starvation
YLR258W	Glycogen synthase; expression induced by glucose limitation, nitrogen starvation, heat shock, and stationary phase; activity regulated by cAMP-dependent, Snf1p and Pho85p kinases as well as by the Gac1p-Glc7p phosphatase; GSY2 has a paralog, GSY1, that arose from the whole genome duplication; relocalizes from cytoplasm to plasma membrane upon DNA replication stress
YLL023C	Transmembrane nucleoporin; involved in nuclear pore complex (NPC) distribution, assembly or stabilization; highly conserved across species, orthologous to human TMEM33 and paralogous to Per33p; protein abundance increases in response to DNA replication stress

Table 7.4: Top 20 indicators common between Ox Gene Expression, Transcription and Histone modifications

Gene ID	Description
YJL192C	ER-membrane protein; subunit of evolutionarily conserved EMC (Endoplasmic Reticulum Membrane Complex) implicated in ERAD (ER-associated degradation) and proper assembly of multi-pass transmembrane (TM) proteins; EMC acts in yeast as an ER-mitochondria tether that interacts with outer membrane protein Tom5 of TOM (Translocase of the Mitochondrial Outer Membrane) complex; suppressor of pma1-7, deletion of SOP4 slows down export of wild-type Pma1p and Pma1-7 from the ER
YKL212W	Phosphatidylinositol phosphate (PtdInsP) phosphatase; involved in hydrolysis of PtdIns[4]P in the early and medial Golgi; regulated by interaction with Vps74p; ER localized transmembrane protein which cycles through the Golgi; involved in protein trafficking and processing, secretion, and cell wall maintenance; regulates sphingolipid biosynthesis through the modulation of PtdIns(4)P metabolism
YDL080C	Regulatory protein that binds Pdc2p and Thi2p transcription factors; activates thiamine biosynthesis transcription factors Pdc2p and Thi2p by binding to them, but releases and de-activates them upon binding to thiamine pyrophosphate (TPP), the end product of the pathway; has similarity to decarboxylases but enzymatic activity is not detected
YOR222W	Mitochondrial inner membrane transporter; 2-oxodicarboxylate transporter, exports 2-oxoadipate and 2-oxoglutarate from the mitochondrial matrix to the cytosol for use in lysine and glutamate biosynthesis and in lysine catabolism; ODC2 has a paralog, ODC1, that arose from the whole genome duplication
YML115C	Component of the mannan polymerase I; complex contains Van1p and Mnn9p and is involved in the first steps of mannan synthesis; mutants are vanadate-resistant
YIL078W	Threonyl-tRNA synthetase; essential cytoplasmic protein; human homolog TARS can complement yeast null mutant
YIL094C	Homo-isocitrate dehydrogenase; an NAD-linked mitochondrial enzyme required for the fourth step in the biosynthesis of lysine, in which homo-isocitrate is oxidatively decarboxylated to alpha-ketoadipate
YPR171W	Adapter that links synaptojanins to the cortical actin cytoskeleton; the synaptojanins are Inp52p and Inp53p
YHR204W	Alpha-1,2-specific exomannosidase of the endoplasmic reticulum; involved in glycan trimming of both folded and misfolded glycoproteins; complexes with Pdi1p, and trims a mannose from Man8GlcNac2 glycans to generate Man7GlcNac2, an oligosaccharide signal on glycoproteins destined for ER-associated protein degradation; requires Pdi1p for stability and substrate recognition; human homolog EDEM1 can complement yeast null mutant
YLR364W	Glutaredoxin that employs a dithiol mechanism of catalysis; monomeric; activity is low and null mutation does not affect sensitivity to oxidative stress; GFP-fusion protein localizes to the cytoplasm; expression strongly induced by arsenic
YMR241W	Citrate and oxoglutarate carrier protein; exports citrate from and imports oxoglutarate into the mitochondrion, causing net export of NADPH reducing equivalents; also associates with mt nucleoids and has a role in replication and segregation of the mt genome
YIL074C	3-phosphoglycerate dehydrogenase and alpha-ketoglutarate reductase; 3PG dehydrogenase that catalyzes the first step in serine and glycine biosynthesis; also functions as an alpha-ketoglutarate reductase, converting alpha-ketoglutarate to D-2-hydroxyglutarate (D-2HG); localizes to the cytoplasm; SER33 has a paralog, SER3, that arose from the whole genome duplication
YPL220W	Ribosomal 60S subunit protein L1A; N-terminally acetylated; homologous to mammalian ribosomal protein L10A and bacterial L1; RPL1A has a paralog, RPL1B, that arose from the whole genome duplication; rpl1a rpl1b double null mutation is lethal
YKL184W	Ornithine decarboxylase; catalyzes the first step in polyamine biosynthesis; degraded in a proteasome-dependent manner in the presence of excess polyamines; deletion decreases lifespan, and increases necrotic cell death and ROS generation
YHR019C	Cytosolic asparaginyl-tRNA synthetase; required for protein synthesis, catalyzes the specific attachment of asparagine to its cognate tRNA
YGR094W	Mitochondrial and cytoplasmic valyl-tRNA synthetase; human homolog VARS2 implicated in mitochondrial diseases, can partially complement yeast null mutant
YIR034C	Saccharopine dehydrogenase (NAD+, L-lysine-forming); catalyzes the conversion of saccharopine to L-lysine, which is the final step in the lysine biosynthesis pathway; also has mRNA binding activity
YPL273W	S-adenosylmethionine-homocysteine methyltransferase; functions along with Mht1p in the conversion of S-adenosylmethionine (AdoMet) to methionine to control the methionine/AdoMet ratio; SAM4 has a paralog, YMR321C, that arose from a single-locus duplication
YMR120C	Enzyme of 'de novo' purine biosynthesis; contains both 5-aminoimidazole-4-carboxamide ribonucleotide transformylase and inosine monophosphate cyclohydrolase activities; ADE17 has a paralog, ADE16, that arose from the whole genome duplication; ade16 ade17 mutants require adenine and histidine
YML055W	Subunit of signal peptidase complex; complex catalyzes cleavage of N-terminal signal sequences of proteins targeted to the secretory pathway; homologous to mammalian SPC25; other members of the complex are Spc1p, Spc1p, and Sec11p

Table 7.5: Rb indicators common between Gene Expression, Transcription and Histone modifications

Gene ID	Description
YGL068W	Mitochondrial ribosomal protein of the large subunit; has similarity to E. coli L7/L12 and human MRPL7 ribosomal proteins; associates with the mitochondrial nucleoid; required for normal respiratory growth
YNL070W	Component of the TOM (translocase of outer membrane) complex; responsible for recognition and initial import steps for all mitochondrially directed proteins; promotes assembly and stability of the TOM complex
YEL050C	Mitochondrial ribosomal protein of the large subunit (L2); has similarity to E. coli L2 ribosomal protein; mutant allele (fat21) causes inability to utilize oleate, and induce oleic acid oxidation; may interfere with activity of the Adr1p transcription factor
YLR253W	Mitochondrial protein of unknown function involved in lipid homeostasis; associates with mitochondrial ribosome; integral membrane protein that localizes to the mitochondrial inner membrane; involved in mitochondrial morphology; non-essential gene which interacts genetically with MDM10, and other members of the ERMES complex; transcription is periodic during the metabolic cycle; homologous to human aarF domain containing kinase, ADCK1
YPR004C	Putative ortholog of mammalian ETF-alpha; interacts with frataxin, Yfh1p; null mutant displays elevated frequency of mitochondrial genome loss; may have a role in oxidative stress response; ETF-alpha is an electron transfer flavoprotein complex subunit
YPL103C	Protein with a role in maintaining mitochondrial morphology; also involved in maintaining normal cardiolipin levels; mitochondrial inner membrane protein; proposed to be involved in N-acylethanolamine metabolism; related to mammalian N-acylPE-specific phospholipase D

Table 7.6: Top 20 indicators common between Rc Gene Expression, Transcription and Histone modifications

Gene ID	Description
YEL060C	Vacuolar proteinase B (yscB) with H3 N-terminal endopeptidase activity; serine protease of the subtilisin family; involved in protein degradation in the vacuole and required for full protein degradation during sporulation; activity inhibited by Pbi2p; protein abundance increases in response to DNA replication stress; PRB1 has a paralog, YSP3, that arose from the whole genome duplication
YCR061W	Protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm in a punctate pattern; induced by treatment with 8-methoxypsoralen and UVA irradiation
YKL100C	Intramembrane aspartyl protease of the perinuclear ER membrane; acts in a branch of ER-associated degradation (ERAD) that degrades functional proteins rather than misfolded proteins; regulates abundance of high-affinity plasma membrane transporters, such as Ctr1p and Zrt1p, during the starvation response; has a presenilin fold; member of the GxGD family of intramembrane proteases; closest human homolog is signal peptide peptidase (SPP)
YJL144W	Cytoplasmic hydrophilin essential in desiccation-rehydration process; expression induced by osmotic stress, starvation and during stationary phase; protein abundance increases in response to DNA replication stress
YLL023C	Transmembrane nucleoporin; involved in nuclear pore complex (NPC) distribution, assembly or stabilization; highly conserved across species, orthologous to human TMEM33 and paralogous to Per33p; protein abundance increases in response to DNA replication stress
YJL141C	Serine-threonine protein kinase; component of a glucose-sensing system that inhibits growth in response to glucose availability; upon nutrient deprivation Yak1p phosphorylates Pop2p to regulate mRNA deadenylation, the co-repressor Crf1p to inhibit transcription of ribosomal genes, and the stress-responsive transcription factors Hsf1p and Msn2p; nuclear localization negatively regulated by the Ras/PKA signaling pathway in the presence of glucose
YJR008W	Protein of unknown function; inhibits haploid invasive growth when overexpressed; synthetically lethal with phospholipase C (PLC1); expression induced by mild heat-stress on a non-fermentable carbon source, upon entry into stationary phase and upon nitrogen deprivation; repressed by inosine and choline in an Opi1p-dependent manner; highly conserved from bacteria to human; Memo, the human homolog, is an ErbB2 interacting protein with an essential function in cell motility
YGR130C	Component of the eisosome with unknown function; GFP-fusion protein localizes to the cytoplasm; specifically phosphorylated in vitro by mammalian diphosphoinositol pentakisphosphate (IP7)
YBR280C	F-Box protein involved in proteasome-dependent degradation of Aah1p; involved in proteasome-dependent degradation of Aah1p during entry of cells into quiescence; interacts with Skp1
YER054C	Putative regulatory subunit of protein phosphatase Glc7p; involved in glycogen metabolism; contains a conserved motif (GVNK motif) that is also found in Gac1p, Pig1p, and Pig2p; GIP2 has a paralog, PIG2, that arose from the whole genome duplication
YNL200C	NADHX epimerase; catalyzes isomerization of (R)- and (S)-NADHX; homologous to AIBP in mammals and the N- terminal domain of YjeF in E.coli; enzyme is widespread in eukaryotes, prokaryotes and archaea; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies
YAL028W	Tail-anchored ER membrane protein of unknown function; interacts with homolog Frt1p; promotes growth in conditions of high Na+, alkaline pH, or cell wall stress, possibly via a role in posttranslational translocation; potential Cdc28p substrate; FRT2 has a paralog, FRT1, that arose from the whole genome duplication
YIR014W	Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the vacuole; expression directly regulated by the metabolic and meiotic transcriptional regulator Ume6p; YIR014W is a non-essential gene
YGR086C	Eisosome core component; eisosomes are large immobile cell cortex structures associated with endocytosis; detected in phosphorylated state in mitochondria; phosphorylated on Thr233 upon Pkc1p hyperactivation in a Slt2p MAPK-dependent fashion; null mutant shows activation of Pkc1p/Ypk1p stress resistance pathways; member of BAR domain family; protein increases in abundance and relocalizes from plasma membrane to cytoplasm upon DNA replication stress
YLR450W	HMG-CoA reductase; converts HMG-CoA to mevalonate, a rate-limiting step in sterol biosynthesis; one of two isozymes; overproduction induces assembly of peripheral ER membrane arrays and short nuclear-associated membrane stacks; forms foci at nuclear periphery upon DNA replication stress; HMG2 has a paralog, HMG1, that arose from the whole genome duplication; human homolog HMGCR can complement yeast hmg2 mutant
YOL032W	Protein with a possible role in phospholipid biosynthesis; null mutant displays an inositol-excreting phenotype that is suppressed by exogenous choline; protein abundance increases in response to DNA replication stress
YBR214W	Protein involved in cell separation during budding; one of two S. cerevisiae homologs (Sds23p and Sds24p) of the S. pombe Sds23 protein, which is implicated in APC/cyclosome regulation; may play an indirect role in fluid-phase endocytosis; protein abundance increases in response to DNA replication stress; SDS24 has a paralog, SDS23, that arose from the whole genome duplication
YAL049C	Cytoplasmic protein involved in mitochondrial function or organization; null mutant displays reduced frequency of mitochondrial genome loss; potential Hsp82p interactor
YDR294C	Dihydrosphingosine phosphate lyase; regulates intracellular levels of sphingolipid long-chain base phosphates (LCBPs), degrades phosphorylated long chain bases, prefers C16 dihydrosphingosine-l-phosphate as a substrate
YKL026C	Phospholipid hydroperoxide glutathione peroxidase; induced by glucose starvation that protects cells from phospholipid hydroperoxides and nonphospholipid peroxides during oxidative stress; GPX1 has a paralog, HYR1, that arose from the whole genome duplication

References

[1] AGUILAR-ARNAL, L. & SASSONE-CORSI, P. (2015). Chromatin landscape and circadian dynamics: Spatial and temporal organization of clock transcription. *Proceedings of the National Academy of Sciences*, **112**, 6863–6870. 111

[2] ANDREWS, S. (2010). FastQC: a quality control tool for high throughput sequence data. 86

[3] ARGELAGUET, R., VELTEN, B., ARNOL, D., DIETRICH, S., ZENZ, T., MARIONI, J.C., BUETTNER, F., HUBER, W. & STEGLE, O. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, **14**, e8124. 30

[4] ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H., CHERRY, J.M., DAVIS, A.P., DOLINSKI, K., DWIGHT, S.S., EPPIG, J.T. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**, 25–29. 57

[5] ATWOOD, A., DECONDE, R., WANG, S.S., MOCKLER, T.C., SABIR, J.S., IDEKER, T. & KAY, S.A. (2011). Cell-autonomous circadian clock of hepatocytes drives rhythms in transcription and polyamine synthesis. *Proceedings of the National Academy of Sciences*, **108**, 18560–18565. 111

[6] BANNISTER, A.J. & KOUZARIDES, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, **21**, 381–395. 4

[7] BARKER, M. & RAYENS, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **17**, 166–173. 27

[8] BAUMGARTNER, U., HAMILTON, B., PISKACEK, M., RUIS, H. & ROTTENSTEINER, H. (1999). Functional Analysis of the Zn 2 Cys 6 Transcription Factors Oaf1p and Pip2p . *Journal of Biological Chemistry*, **274**, 22208–22216. 76

[9] BERGER, S.L. (2002). Histone modifications in transcriptional regulation Berger 143. 142–148. 3, 8, 81

[10] BERGER, S.L. (2007). The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412. 3, 8, 65, 81

[11] BERSANELLI, M., MOSCA, E., REMONDINI, D., GIAMPIERI, E., SALA, C., CASTELLANI, G. & MILANESI, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, **17**, S15. 24

[12] BITTERMAN, K.J., ANDERSON, R.M., COHEN, H.Y., LATORRE-ESTEVES, M. & SINCLAIR, D.A. (2002). Inhibition of silencing and accelerated aging by nicotinamide, a putative negative regulator of yeast Sir2 and human SIRT1. *Journal of Biological Chemistry*, **277**, 45099–45107. 147

[13] BJÖRNE, J. & SALAKOSKI, T. (2013). Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP shared task 2013 workshop*, 16–25. 160

[14] BOLGER, A.M., LOHSE, M. & USADEL, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120. 52, 86

[15] BRO, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, **10**, 47–61. 28, 55

[16] BRO, R. & SMILDE, A.K. (2014). Principal component analysis. *Analytical Methods*, **6**, 2812–2831. 26

[17] BRO, R., SMILDE, A.K. & DE JONG, S. (2001). On the difference between low-rank and subspace approximation: Improved model for multi-linear PLS regression. *Chemometrics and Intelligent Laboratory Systems*, **58**, 3–13. 55, 56

[18] Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, **10**, 1213. 20

[19] Buenrostro, J.D., Wu, B., Chang, H.Y. & Greenleaf, W.J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, **2015**, 21.29.1–21.29.9. 82

[20] Burnetti, A.J., Aydin, M. & Buchler, N.E. (2016). Cell cycle start is coupled to entry into the yeast metabolic cycle across diverse strains and growth rates. *Molecular biology of the cell*, **27**, 64–74. 106, 147

[21] Cai, L. & Tu, B.P. (2011). On acetyl-CoA as a gauge of cellular metabolic state. *Cold Spring Harbor Symposia on Quantitative Biology*, **76**, 195–202. 17, 18, 47, 75, 76, 81, 111, 146, 154, 173, 175, 180

[22] Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J.J. (2016). Transcriptomic and metabolomic data integration. *Briefings in bioinformatics*, **17**, 891–901. 24

[23] Chen, Z.J. & Mas, P. (2019). Interactive roles of chromatin regulation and circadian clock function in plants. *Genome biology*, **20**, 1–12. 111

[24] Churchman, L.S. & Weissman, J.S. (2012). Native elongating transcript sequencing (net-seq). *Current Protocols in Molecular Biology*, **98**, 14–4. 19

[25] Citro, S., Miccolo, C., Meloni, L. & Chiocca, S. (2015). PI3K/mTOR mediate mitogen-dependent HDAC1 phosphorylation in breast cancer: A novel regulation of estrogen receptor expression. *Journal of Molecular Cell Biology*, **7**, 132–142. 175

[26] Conesa, A., Nueda, M.J., Ferrer, A. & Talón, M. (2006). maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102. 53, 85, 87, 92, 122

[27] Conesa, A., Prats-Montalbán, J.M., Tarazona, S., Nueda, M.J. & Ferrer, A. (2010). A multiway approach to data integration in systems biology based on tucker3 and n-pls. *Chemometrics and Intelligent Laboratory Systems*, **104**, 101–111. 27, 28

[28] Conesa, A., Prats-Montalbán, J.M., Tarazona, S., Nueda, M.J. & Ferrer, A. (2010). A multiway approach to data integration in systems biology based on tucker3 and n-pls. *Chemometrics and Intelligent Laboratory Systems*, **104**, 101–111. 102

[29] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. *et al.* (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, **17**, 13. 20, 21

[30] Connelly, L.M. (2011). Cronbach's alpha. *Medsurg nursing*, **20**, 45–47. 118

[31] Consortium, G.O. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, **45**, D331–D338. 57

[32] Consortium, U. (2015). Uniprot: a hub for protein information. *Nucleic acids research*, **43**, D204–D212. 161

[33] Core, L.J., Waterfall, J.J. & Lis, J.T. (2008). Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848. 19

[34] Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. & D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, **42**, 472–477. 29, 153, 154

[35] Cui, X.J., Li, H. & Liu, G.Q. (2011). Combinatorial patterns of histone modifications in saccharomyces. cerevisiae. *Yeast*, **28**, 683–691. 51

[36] Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. & Ashburner, M. (2007). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, **36**, D344–D350. 161

[37] Di Lorenzo, A. & Bedford, M.T. (2011). Histone arginine methylation. *FEBS Letters*, **585**, 2024–2031. 11

[38] Diamantopoulos, A., Siguaw, J.A. & Siguaw, J.A. (2000). *Introducing LISREL: A guide for the uninitiated*. Sage. 25

[39] DIMITROVA, E., TURBERFIELD, A.H. & KLOSE, R.J. (2015). Histone demethylases in chromatin biology and beyond. *EMBO reports*, **16**, 1620–1639. 12

[40] DOHERTY, C.J. & KAY, S.A. (2010). Circadian control of global gene expression patterns. *Annual review of genetics*, **44**, 419–444. 111

[41] DRAPER, N.R. & SMITH, H. (1998). *Applied regression analysis*, vol. 326. John Wiley & Sons. 57

[42] ECKEL-MAHAN, K. & SASSONE-CORSI, P. (2009). Metabolism control by the circadian clock and vice versa. *Nature structural & molecular biology*, **16**, 462–467. 111

[43] ECKEL, R. H., GRUNDY, S. M., & ZIMMET, P.Z. (2005). The metabolic syndrome. *Lancet*, **366**, 1415–1428. 8

[44] ERHARD, F., HALENIUS, A., ZIMMERMANN, C., L'HERNAULT, A., KOWALEWSKI, D.J., WEEKES, M.P., STEVANOVIC, S., ZIMMER, R. & DÖLKEN, L. (2018). Improved ribo-seq enables identification of cryptic translation events. *Nature methods*, **15**, 363–366. 19

[45] ERNST, J. & KELLIS, M. (2017). Chromatin-state discovery and genome annotation with chromhmm. *Nature protocols*, **12**, 2478. 3, 6

[46] ETCHEGARAY, J.P. & MOSTOSLAVSKY, R. (2016). Interplay between Metabolism and Epigenetics: A Nuclear Adaptation to Environmental Changes. *Molecular Cell*, **62**, 695–711. 4, 5, 6, 9, 10, 33

[47] ETCHEGARAY, J.P., LEE, C., WADE, P.A. & REPPERT, S.M. (2003). Rhythmic histone acetylation underlies transcription in the mammalian circadian clock. *Nature*, **421**, 177–182. 13

[48] EVANS, C., HARDIN, J. & STOEBEL, D.M. (2018). Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, **19**, 776–792. 21

[49] FATLAND, B.L., NIKOLAU, B.J. & WURTELE, E.S. (2005). Reverse genetic characterization of cytosolic acetyl-coa generation by atp-citrate lyase in arabidopsis. *The Plant Cell*, **17**, 182–203. 176

[50] FEINGOLD, E.A., GOOD, P.J., GUYER, M.S., KAMHOLZ, S., LIEFER, L., WETTERSTRAND, K., COLLINS, F.S., GINGERAS, T.R., KAMPA, D., SEKINGER, E.A., CHENG, J., HIRSCH, H., GHOSH, S., ZHU, Z., PATEL, S., PICCOLBONI, A., YANG, A., TAMMANA, H., BEKIRANOV, S., KAPRANOV, P., HARRISON, R., CHURCH, G., STRUHL, K., REN, B., KIM, T.H., BARRERA, L.O., QU, C., VAN CALCAR, S., LUNA, R., GLASS, C.K., ROSENFELD, M.G., GUIGO, R., ANTONARAKIS, S.E., BIRNEY, E., BRENT, M., PACHTER, L., REYMOND, A., DERMITZAKIS, E.T., DEWEY, C., KEEFE, D., DENOEUD, F., LAGARDE, J., ASHURST, J., HUBBARD, T., WESSELINK, J.J., CASTELO, R., EYRAS, E., MYERS, R.M., SIDOW, A., BATZOGLOU, S., TRINKLEIN, N.D., HARTMAN, S.J., ALDRED, S.F., ANTON, E., SCHROEDER, D.I., MARTICKE, S.S., NGUYEN, L., SCHMUTZ, J., GRIMWOOD, J., DICKSON, M., COOPER, G.M., STONE, E.A., ASIMENOS, G., BRUDNO, M., DUTTA, A., KARNANI, N., TAYLOR, C.M., KIM, H.K., ROBINS, G., STAMATOYANNOPOULOS, G., STAMATOYANNOPOULOS, J.A., DORSCHNER, M., SABO, P., HAWRYLYCZ, M., HUMBERT, R., WALLACE, J., YU, M., NAVAS, P.A., MCARTHUR, M., NOBLE, W.S., DUNHAM, I., KOCH, C.M., ANDREWS, R.M., CLELLAND, G.K., WILCOX, S., FOWLER, J.C., JAMES, K.D., GROTH, P., DOVEY, O.M., ELLIS, P.D., WRAIGHT, V.L., MUNGALL, A.J., DHAMI, P., FIEGLER, H., LANGFORD, C.F., CARTER, N.P., VETRIE, D., SNYDER, M., EUSKIRCHEN, G., URBAN, A.E., NAGALAKSHMI, U., RINN, J., POPESCU, G., BERTONE, P., HARTMAN, S., ROZOWSKY, J., EMANUELSSON, O., ROYCE, T., CHUNG, S., GERSTEIN, M., LIAN, Z., LIAN, J., NAKAYAMA, Y., WEISSMAN, S., STOLC, V., TONGPRASIT, W., SETHI, H., JONES, S., MARRA, M., SHIN, H., SCHEIN, J., CLAMP, M., LINDBLAD-TOH, K., CHANG, J., JAFFE, D.B., KAMAL, M., LANDER, E.S., MIKKELSEN, T.S., VINSON, J., ZODY, M.C., DE JONG, P.J., OSOEGAWA, K., NEFEDOV, M., ZHU, B., BAXEVANIS, A.D., WOLFSBERG, T.G., CRAWFORD, G.E., WHITTLE, J., HOLT, I.E., VASICEK, T.J., ZHOU, D., LUO, S., GREEN, E.D., BOUFFARD, G.G., MARGULIES, E.H., PORTNOY, M.E., HANSEN, N.F., THOMAS, P.J., MCDOWELL, J.C., MASKERI, B., YOUNG, A.C., IDOL, J.R., BLAKESLEY, R.W., SCHULER, G., MILLER, W., HARDISON, R., ELNITSKI, L., SHAH, P., SALZBERG, S.L., PERTEA, M., MAJOROS, W.H., HAUSSLER, D., THOMAS, D., ROSENBLOOM, K.R., CLAWSON, H., SIEPEL, A., KENT, W.J., WENG, Z., JIN, S., HALEES, A., BURDEN, H., KARAOZ, U., FU, Y., YU, Y., DING, C., CANTOR, C.R., KINGSTON, R.E., DENNIS, J., GREEN, R.D., SINGER, M.A., RICHMOND, T.A., NORTON, J.E., FARNHAM, P.J., OBERLEY, M.J., INMAN, D.R., MCCORMICK, M.R., KIM, H., MIDDLE, C.L., PIRRUNG, M.C., FU, X.D., KWON, Y.S., YE, Z., DEKKER, J., TABUCHI, T.M., GHELDOF, N., DOSTIE, J. & HARVEY, S.C. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, **306**, 636–640. 158

[51] FELTHAM, J., XI, S., MURRAY, S., WOUTERS, M., URDIAIN-ARRAIZA, J., GEORGE, C., TOWNLEY, A., ROBERTS, E., FISHER, R., LIBERATORI, S. *et al.* (2019). Transcriptional changes are regulated by metabolic pathway dynamics but decoupled from protein levels. *bioRxiv*, 833921. 82, 87, 103, 105, 113, 146, 147

[52] FERNÁNDEZ-CID, A., RIERA, A., HERRERO, P. & MORENO, F. (2012). Glucose levels regulate the nucleo-mitochondrial distribution of Mig2. *Mitochondrion*, **12**, 370–380. 76

[53] Finkel, T., Deng, C.X. & Mostoslavsky, R. (2009). Recent progress in the biology and physiology of sirtuins. *Nature*, **460**, 587–591. 10

[54] Forneris, F., Binda, C., Vanoni, M.A., Mattevi, A. & Battaglioli, E. (2005). Histone demethylation catalysed by LSD1 is a flavin-dependent oxidative process. *FEBS Letters*, **579**, 2203–2207. 12

[55] Frank, I.E. & Kowalski, B.R. (1985). A multivariate method for relating groups of measurements connected by a causal pathway. *Analytica Chimica Acta*, **167**, 51–63. 27

[56] Furió-Tarí, P., Conesa, A. & Tarazona, S. (2016). RGmatch: matching genomic regions to proximal genes in omics data integration. *BMC bioinformatics*, **17**, 427. 87, 101

[57] Gardner, C.L., da Silva, D.R., Pagliai, F.A., Pan, L., Padgett-Pagliai, K.A., Blaustein, R.A., Merli, M.L., Zhang, D., Pereira, C., Teplitski, M. *et al.* (2020). Assessment of unconventional antimicrobial compounds for the control of 'candidatus liberibacter asiaticus', the causative agent of citrus greening disease. *Scientific reports*, **10**, 1–15. 166, 177

[58] Gayatri, S. & Bedford, M.T. (2014). Readers of histone methylarginine marks. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1839**, 702–710. 11

[59] Geladi, P. & Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, **185**, 1–17. 26, 54, 124

[60] Goh, G.H., Maloney, S.K., Mark, P.J. & Blache, D. (2019). Episodic ultradian events-ultradian rhythms. *Biology*, **8**. 12

[61] Gowans, G.J., Schep, A.N., Wong, K.M., King, D.A., Greenleaf, W.J. & Morrison, A.J. (2018). INO80 Chromatin Remodeling Coordinates Metabolic Homeostasis with Cell Division. *Cell Reports*, **22**, 611–623. 107, 111, 148

[62] Grant, C.E., Bailey, T.L. & Noble, W.S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018. 87, 101

[63] Grüne, T., Brzeski, J., Eberharter, A., Clapier, C.R., Corona, D.F., Becker, P.B. & Müller, C.W. (2003). Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor iswi. *Molecular cell*, **12**, 449–460. 7

[64] Gut, P. & Verdin, E. (2013). The nexus of chromatin regulation and intermediary metabolism. *Nature*, **502**, 489. 8

[65] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr, M., Jungkamp, A.C., Munschauer, M. *et al.* (2010). Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, **141**, 129–141. 19

[66] Hamann, T. (2015). The plant cell wall integrity maintenance mechanism—concepts for organization and mode of action. *Plant and Cell Physiology*, **56**, 215–223. 176

[67] Hansen, K.D., Irizarry, R.A. & Wu, Z. (2012). Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216. 21

[68] Harvey, A., Caretti, G., Moresi, V., Renzini, A. & Adamo, S. (2019). Interplay between metabolites and the epigenome in regulating embryonic and adult stem cell potency and maintenance. *Stem cell reports*, **13**, 573–589. 12

[69] Hasanuzzaman, M., Nahar, K., Hossain, M., Mahmud, J.A., Rahman, A., Inafuku, M., Oku, H., Fujita, M. *et al.* (2017). Coordinated actions of glyoxalase and antioxidant defense systems in conferring abiotic stress tolerance in plants. *International journal of molecular sciences*, **18**, 200. 176

[70] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A. *et al.* (2004). Intact: an open source molecular interaction database. *Nucleic acids research*, **32**, D452–D455. 29

[71] Hernández-de Diego, R., Boix-Chova, N., Gómez-Cabrero, D., Tegner, J., Abugessaisa, I. & Conesa, A. (2014). STATegra EMS: an Experiment Management System for complex next-generation omics experiments. *BMC systems biology*, **8**, S9. 24

[72] Hernández-De-Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas, G.J. & Conesa, A. (2018). PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Research*, **46**, W503–W509. 29, 30

[73] Hobert, O. (2004). Common logic of transcription factor and microRNA action. *Trends in Biochemical Sciences*, **29**, 462–468. 3

[74] Howe, F.S., Fischl, H., Murray, S.C. & Mellor, J. (2017). Is h3k4me3 instructive for transcription activation? *Bioessays*, **39**, 1–12. 76

[75] Huang, S., Chaudhary, K. & Garmire, L.X. (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, **8**, 84. 24

[76] Hughes, M.E., Hong, H.K., Chong, J.L., Indacochea, A.A., Lee, S.S., Han, M., Takahashi, J.S. & Hogenesch, J.B. (2012). Brain-specific rescue of clock reveals system-driven transcriptional rhythms in peripheral tissue. *PLoS Genet*, **8**, e1002835. 111

[77] ichiro Imai, S. & Guarente, L. (2014). NAD+ and sirtuins in aging and disease. *Trends in Cell Biology*, **24**, 464–471. 9

[78] Jenuwein, T. & Allis, C.D. (2001). Translating the histone code. *Science*, **293**, 1074–1080. 9

[79] Julius, B.T., Leach, K.A., Tran, T.M., Mertz, R.A. & Braun, D.M. (2017). Sugar transporters in plants: new insights and discoveries. *Plant and Cell Physiology*, **58**, 1442–1460. 176

[80] Kaelin, W.G. & McKnight, S.L. (2013). Influence of metabolism on epigenetics and disease. *Cell*, **153**, 56–69. 12

[81] Kanani, H., Chrysanthopoulos, P.K. & Klapa, M.I. (2008). Standardizing gc–ms metabolomics. *Journal of Chromatography B*, **871**, 191–201. 21

[82] Kandasamy, K., Mohan, S.S., Raju, R., Keerthikumar, S., Kumar, G.S.S., Venugopal, A.K., Telikicherla, D., Navarro, J.D., Mathivanan, S., Pecquet, C. *et al.* (2010). Netpath: a public resource of curated signal transduction pathways. *Genome biology*, **11**, 1–9. 29

[83] Kanehisa, M. & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27–30. 57

[84] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, **44**, D457–D462. 57

[85] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, **45**, D353–D361. 28, 153

[86] Kankainen, M., Gopalacharyulu, P., Holm, L. & Orešič, M. (2011). Mpea—metabolite pathway enrichment analysis. *Bioinformatics*, **27**, 1878–1879. 29

[87] Karp, P.D., Riley, M., Paley, S.M. & Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic acids research*, **30**, 59–61. 29, 153, 165

[88] Karpichev, I.V., Luo, Y., Marians, R.C. & Small, G.M. (1997). A complex containing two transcription factors regulates peroxisome proliferation and the coordinate induction of beta-oxidation enzymes in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, **17**, 69–80. 76

[89] Karpievitch, Y.V., Taverner, T., Adkins, J.N., Callister, S.J., Anderson, G.A., Smith, R.D. & Dabney, A.R. (2009). Normalization of peak intensities in bottom-up ms-based proteomics using singular value decomposition. *Bioinformatics*, **25**, 2573–2580. 22

[90] Katada, S. & Sassone-Corsi, P. (2010). The histone methyltransferase mll1 permits the oscillation of circadian gene expression. *Nature structural & molecular biology*, **17**, 1414. 111

[91] Kelder, T., Van Iersel, M.P., Hanspers, K., Kutmon, M., Conklin, B.R., Evelo, C.T. & Pico, A.R. (2012). WikiPathways: Building research communities on biological pathways. *Nucleic Acids Research*, **40**, 1301–1307. 154

[92] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U. *et al.* (2012). The intact molecular interaction database in 2012. *Nucleic acids research*, **40**, D841–D846. 158

[93] Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. & Karp, P.D. (2005). EcoCyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, **33**, 334–337. 171

[94] Kettner, N.M., Mayo, S.A., Hua, J., Lee, C., Moore, D.D. & Fu, L. (2015). Circadian dysfunction induces leptin resistance in mice. *Cell Metabolism*, **22**, 448–459. 13

[95] KIM, J.D., OHTA, T., PYYSALO, S., KANO, Y. & TSUJII, J. (2009). Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 workshop companion volume for shared task*, 1–9. 154, 160

[96] KLEMM, S.L., SHIPONY, Z. & GREENLEAF, W.J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, **20**, 207–220. 3

[97] KLEVECZ, R.R., BOLEN, J., FORREST, G. & MURRAY, D.B. (2004). A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 1200–1205. 47, 74, 111

[98] KÖNIG, J., ZARNACK, K., LUSCOMBE, N.M. & ULE, J. (2012). Protein–rna interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, **13**, 77–83. 19

[99] KORNMANN, B., SCHAAD, O., BUJARD, H., TAKAHASHI, J.S. & SCHIBLER, U. (2007). System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS Biol*, **5**, e34. 111

[100] KRALLINGER, M., MORGAN, A., SMITH, L., LEITNER, F., TANABE, L., WILBUR, J., HIRSCHMAN, L. & VALENCIA, A. (2008). Evaluation of text-mining systems for biology: Overview of the Second BioCreative community challenge. *Genome Biology*, **9**, 1–9. 158

[101] KRAMER, O. (2016). Benchmark Functions. *Studies in Big Data: Machine Learning for Evolution Strategies*, **20**, 119–124. 165

[102] 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 and Molecular Biology*, **21**, 854–863. 15, 17, 47, 48, 51, 59, 61, 74, 75, 111, 113, 124, 146

[103] KUANG, Z., JI, Z., BOEKE, J.D. & JI, H. (2018). Dynamic motif occupancy (DynaMO) analysis identifies transcription factors and their binding sites driving dynamic biological processes. *Nucleic acids research*, **46**, e2. 17, 48, 111

[104] KURDISTANI, S.K. & GRUNSTEIN, M. (2003). Histone acetylation and deacetylation in yeast. *Nature Reviews Molecular Cell Biology*, **4**, 276–284. 9

[105] LANGMEAD, B. & SALZBERG, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357. 86

[106] LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**. 52

[107] LÄNGST, G. & MANELYTE, L. (2015). Chromatin remodelers: From function to dysfunction. *Genes*, **6**, 299–324. 7

[108] LAW, C.W., CHEN, Y., SHI, W. & SMYTH, G.K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, **15**, R29. 23

[109] LEAMAN, R., WEI, C.H. & LU, Z. (2015). TmChem: A high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, **7**, 1–10. 179

[110] LEAMAN, ROBERT AND GONZALEZ, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*, 652—-663, World Scientific. 179

[111] LEE, C.Y. & GRANT, P.A. (2018). *Role of histone acetylation and acetyltransferases in gene regulation*. Elsevier Inc. 8, 9

[112] LEE, K.K. & WORKMAN, J.L. (2007). Histone acetyltransferase complexes: One size doesn't fit all. *Nature Reviews Molecular Cell Biology*, **8**, 284–295. 6, 8

[113] LEEK, J.T., JOHNSON, W.E., PARKER, H.S., JAFFE, A.E. & STOREY, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883. 22, 87

[114] LEMOINE, R., LA CAMERA, S., ATANASSOVA, R., DÉDALDÉCHAMP, F., ALLARIO, T., POURTAU, N., BONNEMAIN, J.L., LALOI, M., COUTOS-THÉVENOT, P., MAUROUSSET, L. *et al.* (2013). Source-to-sink transport of sugar and regulation by environmental factors. *Frontiers in plant science*, **4**, 272. 176

[115] LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. 52

[116] LI, Q., ZHOU, H., WURTELE, H., DAVIES, B., HORAZDOVSKY, B., VERREAULT, A. & ZHANG, Z. (2008). Acetylation of histone h3 lysine 56 regulates replication-coupled nucleosome assembly. *Cell*, **134**, 244–255. 66, 76

[117] LI, W., ZHANG, S., LIU, C.C. & ZHOU, X.J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466. 27

[118] LIAO, Y., SMYTH, G.K. & SHI, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930. 87

[119] LOVE, M., ANDERS, S. & HUBER, W. (2014). Differential analysis of count data–the deseq2 package. *Genome Biol*, **15**, 10–1186. 23

[120] LU, C. & THOMPSON, C.B. (2012). Metabolic regulation of epigenetics. *Cell Metabolism*, **16**, 9–17. 150

[121] MARDIS, E.R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry*, **6**, 287–303. 19

[122] MARTINS, C.D.P.S., PEDROSA, A.M., DU, D., GONCALVES, L.P., YU, Q., GMITTER JR, F.G. & COSTA, M.G.C. (2015). Genome-wide characterization and expression analysis of major intrinsic proteins during abiotic and biotic stresses in sweet orange (citrus sinensis l. osb.). *PLoS one*, **10**, e0138786. 176

[123] MASRI, S. & SASSONE-CORSI, P. (2014). Sirtuins and the circadian clock: Bridging chromatin and metabolism. *Science Signaling*, **7**, 1–7. 13

[124] MASUI, K., TANAKA, K., AKHAVAN, D., BABIC, I., GINI, B., MATSUTANI, T., IWANAMI, A., LIU, F., VILLA, G.R., GU, Y., CAMPOS, C., ZHU, S., YANG, H., YONG, W.H., CLOUGHESY, T.F., MELLINGHOFF, I.K., CAVENEE, W.K., SHAW, R.J. & MISCHEL, P.S. (2013). MTOR complex 2 controls glycolytic metabolism in glioblastoma through FoxO acetylation and upregulation of c-Myc. *Cell Metabolism*, **18**, 726–739. 175

[125] MEIER, L., VAN DE GEER, S. & BÜHLMANN, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 53–71. 25

[126] MELLOR, J. (2016). The molecular basis of metabolic cycles and their relationship to circadian rhythms. *Nature Structural and Molecular Biology*, **23**, 1035–1044. 15, 47, 75, 81, 111, 146

[127] MENG, C., KUSTER, B., CULHANE, A.C. & GHOLAMI, A.M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, **15**, 162. 26

[128] MENTCH, S.J., MEHRMOHAMADI, M., HUANG, L., LIU, X., GUPTA, D., MATTOCKS, D., PADILLA, P.G., ABLES, G., BAMMAN, M.M. & THALACKER-MERCER, A.E. (2015). Histone methylation dynamics and gene regulation occur through the sensing of one-carbon metabolism. *Cell metabolism*, **22**, 861–873. 11, 77

[129] MEWS, P., DONAHUE, G., DRAKE, A.M., LUCZAK, V., ABEL, T. & BERGER, S.L. (2017). Acetyl-CoA synthetase regulates histone acetylation and hippocampal memory. *Nature*, **546**, 381–386. 154

[130] MILLER, J.J. (2013). Graph database applications and concepts with Neo4j. *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, **2324**, 36. 155, 161, 179

[131] MISHUR, R.J. & REA, S.L. (2012). Applications of mass spectrometry to metabolomics and metabonomics: Detection of biomarkers of aging and of age-related diseases. *Mass spectrometry reviews*, **31**, 70–95. 21

[132] MOHLER, R.E., TU, B.P., DOMBEK, K.M., HOGGARD, J.C., YOUNG, E.T. & SYNOVEC, R.E. (2008). Identification and evaluation of cycling yeast metabolites in two-dimensional comprehensive gas chromatography-time-of-flight-mass spectrometry data. *Journal of Chromatography A*, **1186**, 401–411. 47

[133] MORTAZAVI, A., WILLIAMS, B.A., MCCUE, K., SCHAEFFER, L. & WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**, 621–628. 21

[134] MURRAY, D.B., BECKMANN, M. & KITANO, H. (2007). Regulation of yeast oscillatory dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2241–2246. 47

[135] NAKAHATA, Y., SAHAR, S., ASTARITA, G., KALUZOVA, M. & SASSONE-CORSI, P. (2009). Circadian Control of the NAD+ Salvage Pathway by CLOCK-SIRT1. *Science*, **324**, 654–657. 13

[136] NELDER, J.A. & WEDDERBURN, R.W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**, 370–384. 24

[137] NUEDA, M.J., FERRER, A. & CONESA, A. (2012). Arsyn: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics*, **13**, 553–566. 22

[138] NUEDA, M.J., TARAZONA, S. & CONESA, A. (2014). Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, **30**, 2598–2602. 23, 53, 85, 87, 92, 113

[139] O'BRIEN, K.P., REMM, M. & SONNHAMMER, E.L. (2005). Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, **33**, 476–480. 164

[140] PARK, J.M., KIM, T.H., JO, S.H., KIM, M.Y. & AHN, Y.H. (2015). Acetylation of glucokinase regulatory protein decreases glucose metabolism by suppressing glucokinase activity. *Scientific Reports*, **5**, 1–13. 173

[141] PARK, P.J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680. 19, 81

[142] PATUMCHAROENPOL, P., DOUNGPAN, N. & MEECHAI, A. (2016). An integrated text mining framework for metabolic interaction network reconstruction. 1–23. 160

[143] PAVKOVIC, M., PANTANO, L., GERLACH, C.V., BRUTUS, S., BOSWELL, S.A., EVERLEY, R.A., SHAH, J.V., SUI, S.H. & VAIDYA, V.S. (2019). Multi omics analysis of fibrotic kidneys in two mouse models. *Scientific data*, **6**, 1–9. 24

[144] PINU, F.R., BEALE, D.J., PATEN, A.M., KOUREMENOS, K., SWARUP, S., SCHIRRA, H.J. & WISHART, D. (2019). Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites*, **9**, 1–31. 121

[145] PIPER, J., ELZE, M.C., CAUCHY, P., COCKERILL, P.N., BONIFER, C. & OTT, S. (2013). Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, **41**. 87, 100

[146] PORTALES-CASAMAR, E., ARENILLAS, D., LIM, J., SWANSON, M.I., JIANG, S., MCCALLUM, A., KIROV, S. & WASSERMAN, W.W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, **37**, 54–60. 29, 153, 158

[147] QUINLAN, A.R. & HALL, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842. 52

[148] RAI, V. (2002). Role of amino acids in plant responses to stresses. *Biologia plantarum*, **45**, 481–487. 176

[149] RAMSEY, K.M., YOSHINO, J., BRACE, C.S., ABRASSART, D., KOBAYASHI, Y., MARCHEVA, B., HONG, H.K., CHONG, J.L., BUHR, E.D., LEE, C., TAKAHASHI, J.S., IMAI, S.I. & BASS, J. (2009). Circadian clock feedback cycle through NAMPT-Mediated NAD+ biosynthesis. *Science*, **324**, 651–654. 13

[150] RAO, A.R. & PELLEGRINI, M. (2011). Regulation of the yeast metabolic cycle by transcription factors with periodic activities. *BMC Systems Biology*, **5**. 48

[151] REIMAND, J., ISSERLIN, R., VOISIN, V., KUCERA, M., TANNUS-LOPES, C., ROSTAMIANFAR, A., WADI, L., MEYER, M., WONG, J., XU, C. *et al.* (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, **14**, 482–517. 29

[152] REPPERT, S.M. & WEAVER, D.R. (2002). Coordination of circadian clocks in mammals. *Nature*, **418**, 935–941. 12

[153] ROBINSON, M.D. & OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**, 1–9. 22

[154] ROBINSON, M.D., MCCARTHY, D.J. & SMYTH, G.K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140. 23

[155] ROHART, F., GAUTIER, B., SINGH, A. & LÊ CAO, K.A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, **13**, e1005752. 26

[156] ROUSSEEUW, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65. 54

[157] ROY, S. & TENNISWOOD, M. (2007). Site-specific acetylation of p53 directs selective transcription complex assembly. *Journal of Biological Chemistry*, **282**, 4765–4771. 175

[158] RUDIC, R.D., MCNAMARA, P., CURTIS, A.M., BOSTON, R.C., PANDA, S., HOGENESCH, J.B. & FITZGERALD, G.A. (2004). Bmal1 and clock, two essential components of the circadian clock, are involved in glucose homeostasis. *PLoS Biol*, **2**, e377. 111

[159] S. Wold, H. Martens, H.W. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *The Matrix Pencil*, 286–293. 54

[160] Sanchez, G. (2013). PLS Path Modeling with R. *R Package Notes*, 235. 116

[161] Sanchez, G., Trinchera, L. & Russolillo, G. (2017). plspm: Tools for Partial Least Squares Path Modeling (PLSPM). *R package version 0.4.9*. 117, 118

[162] Schomburg, I. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, **32**, 431D–433. 158

[163] Shahbazian, M.D. & Grunstein, M. (2007). Functions of Site-Specific histone acetylation and deacetylation. *Annual Review of Biochemistry*, **76**, 75–100. 6

[164] Shao, W. & Espenshade, P.J. (2012). Expanding roles for srebp in metabolism. *Cell metabolism*, **16**, 414–419. 175

[165] Shi, X., Liu, S., Metges, C.C. & Seyfert, H.M. (2010). C/EBP-beta drives expression of the nutritionally regulated promoter IA of the acetyl-CoA carboxylase-alpha gene in cattle. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1799**, 561–567. 175

[166] Simõ-Riudalbas, L. & Esteller, M. (2015). Targeting the histone orthography of cancer: Drugs for writers, erasers and readers. *British Journal of Pharmacology*, **172**, 2716–2732. 8

[167] Slavov, N., Macinskas, J., Caudy, A. & Botstein, D. (2011). Metabolic cycling without cell division cycling in respiring yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 19090–19095. 47, 74

[168] Smilde, A., Bro, R. & Geladi, P. (2005). *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons. 27

[169] Smilde, A.K., Jansen, J.J., Hoefsloot, H.C., Lamers, R.J.A., Van Der Greef, J. & Timmerman, M.E. (2005). Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048. 22

[170] Sone, H., Shimano, H., Sakakura, Y., Inoue, N., Amemiya-kudo, M., Yahagi, N., Osawa, M., Suzuki, H., Yokoo, T., Takahashi, A., Iida, K., Toyoshima, H., Iwama, A., Yamada, N., Shimano, H., Sakakura, Y., Inoue, N., Amemiya-kudo, M., Yahagi, N., Osawa, M., Suzuki, H., Yokoo, T., Takahashi, A., Iida, K., Toyoshima, H., Iwama, A. & Acetyl-coenzyme, N.Y. (2020). Acetyl-coenzyme A synthetase is a lipogenic enzyme controlled by SREBP-1 and energy status. 222–230. 175

[171] Soto, A.J., Zerva, C., Batista-Navarro, R. & Ananiadou, S. (2018). LitPathExplorer: A confidence-based visual text analytics tool for exploring literature-enriched pathway models. *Bioinformatics*, **34**, 1389–1397. 154

[172] Stevenson, T.J. (2018). Epigenetic Regulation of Biological Rhythms: An Evolutionary Ancient Molecular Timer. *Trends in Genetics*, **34**, 90–100. 146

[173] Swainston, N., Batista-Navarro, R., Carbonell, P., Dobson, P.D., Dunstan, M., Jervis, A.J., Vinaixa, M., Williams, A.R., Ananiadou, S., Faulon, J.L., Mendes, P., Kell, D.B., Scrutton, N.S. & Breitling, R. (2017). biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS ONE*, **12**, 1–14. 154

[174] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J. & Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, **45**, D362–D368. 29, 153, 158

[175] Takahashi, J.S. (2017). Transcriptional architecture of the mammalian circadian clock. *Nature Reviews Genetics*, **18**, 164–179. 111

[176] Takahashi, J.S. (2017). Transcriptional architecture of the mammalian circadian clock. *Nature Reviews Genetics*, **18**, 164. 111

[177] Takatsuji, H. & Jiang, C.J. (2014). Plant hormone crosstalks under biotic stresses. In *Phytohormones: a window to metabolism, signaling and biotechnological applications*, 323–350, Springer. 176

[178] Takusagawa, F., Kamitori, S., Misaki, S. & Markham, G.D. (1996). Crystal structure of S-adenosylmethionine synthetase. *Journal of Biological Chemistry*, **271**, 136–147. 11

[179] Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, J., Ye, Y., Khochbin, S., Ren, B. & Zhao, Y. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, **146**, 1016–1028. 5

[180] Tarazona, S., García, F., Ferrer, A., Dopazo, J. & Conesa, A. (2011). Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, **17**, 18–19. 23

[181] Tarazona, S., Balzano-Nogueira, L. & Conesa, A. (2018). *Multiomics Data Integration in Time Series Experiments*, vol. 82. Elsevier B.V., 1st edn. 24, 25, 34, 81, 125

[182] Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., Tegnér, J., Westerhuis, J.A. & Conesa, A. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nature communications*, **11**, 1–13. 23

[183] Teixeira, M.C., Monteiro, P.T., Palma, M., Costa, C., Godinho, C.P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T. & Sá-Correia, I. (2018). YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Research*, **46**, D348–D353. 57, 69

[184] Thévenot, E.A., Roux, A., Xu, Y., Ezan, E. & Junot, C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and opls statistical analyses. *Journal of proteome research*, **14**, 3322–3335. 26

[185] Tsuruoka, Y., McNaught, J. & Ananiadou, S. (2008). Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, **9**, 1–10. 163

[186] Tu, B.P. & McKnight, S.L. (2009). Evidence of carbon monoxide-mediated phase advancement of the yeast metabolic cycle. *Proceedings of the National Academy of Sciences*, **106**, 14293–14296. 18

[187] Tu, B.P., Kudlicki, A., Rowicka, M. & McKnight, S.L. (2005). Cell biology: Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, **310**, 1152–1158. 15, 17, 47, 51, 61, 74, 111, 146, 147

[188] Tu, B.P., Mohler, R.E., Liu, J.C., Dombek, K.M., Young, E.T., Synovec, R.E. & McKnight, S.L. (2007). Cyclic changes in metabolic state during the life of a yeast cell. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 16886–16891. 18, 47, 81, 111

[189] Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, **13**, 966–967. 29, 153, 158

[190] Tyagi, M., Imam, N., Verma, K. & Patel, A.K. (2016). Chromatin remodelers: We are the drivers!! *Nucleus*, **7**, 388–404. 7, 8

[191] Ullman, J.B. & Bentler, P.M. (2003). Structural equation modeling. *Handbook of psychology*, 607–634. 25

[192] Upchurch, R.G. (2008). Fatty acid unsaturation, mobilization, and regulation in the response of plants to stress. *Biotechnology letters*, **30**, 967–977. 176

[193] van der Kloet, F.M., Sebastián-León, P., Conesa, A., Smilde, A.K. & Westerhuis, J.A. (2016). Separating common from distinctive variation. *BMC bioinformatics*, **17**, S195. 26

[194] Wang, X.Z. & Ron, D. (1996). Stress-induced phosphorylation and activation of the transcription factor CHOP (GADD153) by p38 MAP kinase. *Science*, **272**, 1347–1349. 175

[195] Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M. & Leser, U. (2020). Huner: improving biomedical ner with pretraining. *Bioinformatics*, **36**, 295–302. 180

[196] Weiner, A., Hsieh, T.H.S., Appleboim, A., Chen, H.V., Rahat, A., Amit, I., Rando, O.J. & Friedman, N. (2015). High-resolution chromatin dynamics during a yeast stress response. *Molecular Cell*, **58**, 371–386. 51

[197] Wellen, K.E. & Thompson, C.B. (2012). A two-way street: Reciprocal regulation of metabolism and signalling. *Nature Reviews Molecular Cell Biology*, **13**, 270–276. 9, 10, 175, 180

[198] Wellen, K.E., Hatzivassiliou, G., Sachdeva, U.M., Bui, T.V., Cross, J.R. & Thompson, C.B. (2009). ATP-citrate lyase links cellular metabolism to histone acetylation. *Science*, **324**, 1076–1080. 6, 8, 9, 33, 75, 81, 146, 154, 173

[199] WIERMAN, M.B. & SMITH, J.S. (2014). Yeast sirtuins and the regulation of aging. *FEMS Yeast Research*, **14**, 73–88. 10

[200] WINGENDER, E., DIETZE, P., KARAS, H. & KNÜPPEL, R. (1996). Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, **24**, 238–241. 29

[201] WOLD, H. (1975). Path models with latent variables: The nipals approach. In *Quantitative sociology*, 307–357, Elsevier. 28

[202] WU, G.A., TEROL, J., IBANEZ, V., LÓPEZ-GARCÍA, A., PÉREZ-ROMÁN, E., BORREDÁ, C., DOMINGO, C., TADEO, F.R., CARBONELL-CABALLERO, J., ALONSO, R. *et al.* (2018). Genomics of the origin and evolution of citrus. *Nature*, **554**, 311–316. 175

[203] WU, Y. & LI, L. (2016). Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A*, **1430**, 80–95. 22

[204] YOU, H. & MAK, T.W. (2005). Crosstalk between p53 and foxo transcription factors. *Cell Cycle*, **4**, 37–38. 175

[205] YU, E.Y., STEINBERG-NEIFACH, O., DANDJINOU, A.T., KANG, F., MORRISON, A.J., SHEN, X. & LUE, N.F. (2007). Regulation of telomere structure and functions by subunits of the ino80 chromatin remodeling complex. *Molecular and cellular biology*, **27**, 5639–5649. 7

[206] ZERBINO, D.R., ACHUTHAN, P., AKANNI, W., AMODE, M.R., BARRELL, D., BHAI, J., BILLIS, K., CUMMINS, C., GALL, A., GIRÓN, C.G., GIL, L., GORDON, L., HAGGERTY, L., HASKELL, E., HOURLIER, T., IZUOGU, O.G., JANACEK, S.H., JUETTEMANN, T., TO, J.K., LAIRD, M.R., LAVIDAS, I., LIU, Z., LOVELAND, J.E., MAUREL, T., MCLAREN, W., MOORE, B., MUDGE, J., MURPHY, D.N., NEWMAN, V., NUHN, M., OGEH, D., ONG, C.K., PARKER, A., PATRICIO, M., RIAT, H.S., SCHUILENBURG, H., SHEPPARD, D., SPARROW, H., TAYLOR, K., THORMANN, A., VULLO, A., WALTS, B., ZADISSA, A., FRANKISH, A., HUNT, S.E., KOSTADIMA, M., LANGRIDGE, N., MARTIN, F.J., MUFFATO, M., PERRY, E., RUFFIER, M., STAINES, D.M., TREVANION, S.J., AKEN, B.L., CUNNINGHAM, F., YATES, A. & FLICEK, P. (2018). Ensembl 2018. *Nucleic Acids Research*, **46**, D754–D761. 52

[207] ZHANG, E.E., LIU, Y., DENTIN, R., PONGSAWAKUL, P.Y., LIU, A.C., HIROTA, T., NUSINOW, D.A., SUN, X., LANDAIS, S., KODAMA, Y. *et al.* (2010). Cryptochrome mediates circadian regulation of camp signaling and hepatic gluconeogenesis. *Nature medicine*, **16**, 1152. 111

[208] ZHANG, P., DREHER, K., KARTHIKEYAN, A., CHI, A., PUJAR, A., CASPI, R., KARP, P., KIRKUP, V., LATENDRESSE, M., LEE, C., MUELLER, L.A., MULLER, R. & RHEE, S.Y. (2010). Creation of a genome-wide metabolic pathway database for populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiology*, **153**, 1479–1491. 176

[209] ZHANG, Y., LIU, T., MEYER, C.A., EECKHOUTE, J., JOHNSON, D.S., BERNSTEIN, B.E., NUSBAUM, C., MYERS, R.M., BROWN, M. & LI, W. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**, R137. 20, 87

[210] ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, **67**, 301–320. 25

[211] ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **67**, 301–320. 57

Vniversitat de València