Modelización del Ciclo Metabólico de la Levadura mediante la Integración Estadística de Datos de Metabolómica, Expresión Génica y Modificación de Histonas.

Sergio Doria Belenguer
Máster en Bioinformática y Bioestadística
Área 3: Anotación e Integración de Datos Ómicos

Nombre Consultor/a: Sonia Tarazona Campos
Nombre Profesor/a responsable de la asignatura Ferran Prados Carrasco

04/01/2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Modelización del Ciclo Metabólico de la Levadura mediante la Integración Estadística de Datos de Metabolómica, expresión génica y modificación de histonas.*
Nombre del autor:	*Sergio Doria Belenguer*
Nombre del consultor/a:	*Sonia Tarazona Campos*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	01/2019
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 3: Anotación e Integración de Datos Ómicos*
Idioma del trabajo:	*Español*
Palabras clave	*Biología de sistemas, Integración estadística, datos ómicos*

Resumen del Trabajo:

En los últimos años la biología de sistemas se ha posicionado como una de las áreas más interesantes en la investigación biológica. La comprensión de la célula como un sistema integrado promete el descubrimiento de nuevas propiedades emergentes nunca antes descritas. Sin embargo, aunque el avance de las tecnologías ómicas ha permitido la generación de grandes cantidades de datos, su integración e interpretación siguen siendo un reto a superar.

En este contexto uno de los modelos más estudiados es el ciclo metabólico de la levadura (YMC, de sus siglas en inglés). Este ciclo aparece en situaciones de limitación de nutrientes y se caracteriza por una oscilación periódica de la expresión génica. Recientes estudios sugieren que dicha oscilación es producto de cambios en el metaboloma que podrían afectar a la estructura de la cromatina. Esto hace al YMC un modelo perfecto para el análisis de la relación entre metaboloma, transcriptoma y estado de la cromatina.

Con el objetivo de desarrollar nuevos protocolos de integración estadística de datos ómicos, se utilizaron datos de RNA-Seq, ChIP-Seq y metabolómica del YMC. Para ello se emplearon diferentes modelos estadísticos: modelos multivariantes (PLS) y de regresión lineal múltiple (MORE). Tras su aplicación se encontraron diferencias entre ambos modelos. Finalmente se realizó una interpretación biológica de los resultados.

Abstract:

In the last years system biology has become in one of the most exciting fields in biological research. The understanding of the cell as an interactive network, instead of separated layers, could be decisive in the comprehension of human diseases and development. This is because the interaction analyses may uncover emerging proprieties that lie hidden. Despite nowadays technology allows researchers to produce large amount of data from different omic technology; the statistical integration and interpretation is not easy and available tools are far to be perfect. Consequently, the ideal system model is not achieved yet and more efforts are needed.

In this context, one of the most studied models is the yeast metabolic cycle (YMC). This cycle is defined by a periodic oscillation of genetic expression. Based on recent studies, during this cycle the regulation of the genes could be done by metabolic changes which affects the chromatin structure. This makes the YMC a perfect model to study the relation between metabolomic, transcriptomic and ChIP-seq data in an integrative way.

The objective of this research is the contribution to the development of a statistical and computational pipeline for the integration of RNA-Seq, ChIP-Seq and metabolomic data from YMC. To achieve it different statistical approaches were used: multivariate models (PLS) and multiple lineal regression models (MORE). Clear differences between models were found and discussed. Finally, a biological interpretation of the results was done.

AGRADECIMIENTOS

Primero de todo me gustaría agradecer este trabajo a todo el equipo del laboratorio i52 del CIPF por hacerme sentir parte de la familia y ayudarme en todo momento. Gracias a Carlos por enseñarme los secretos del *apply* y a programar artísticamente. A Manu por sus consejos estadísticos y sus chistes, no siempre políticamente correctos. También a Teresa y Pedro por ayudarme a desconectar los viernes dejando un ordenador para coger otro. A Ángeles, siempre dispuesta a generar discusiones interesantes. A Fran, al que nunca vi llegar a la hora. A Cristina, la abuelilla del lab siempre acompañada de su "motoreta". Como olvidarme de Salva, siempre preocupado por la mejora de todos en las presentaciones en inglés del temido *Journal Club*. Y, finalmente a mi tutora Sonia, siempre dispuesta a ayudar con una sonrisa, para ella solo tengo buenas palabras, esta tesis no habría sido posible sin ella.

Índice

1. Introducción 1
1.1 Contexto y justificación del Trabajo 1
1.2 Objetivos del Trabajo 2
1.3 Enfoque y método seguido 2
1.4 Planificación del Trabajo 4
1.4.1 Riesgos y puntos clave en el trabajo 7
1.5 Breve sumario de productos obtenidos 8
1.6 Breve descripción de los otros capítulos de la memoria 8
2. Materiales y Métodos 9
2.1 Datos Ómicos 9
2.1.1 Datos RNA-Seq 9
2.1.2 Datos ChIP-Seq 10
2.1.3 Datos Metabolómicos 10
2.2 Preprocesado de los Datos de Metabolómica 11
2.2.1 Análisis Exploratorio 11
2.2.2 Construcción de la Matriz Final de Datos de Metabolómica 14
2.2.3 Datos Faltantes 14
2.3 Expresión Diferencial (maSigPro) 19
2.4 Regulaciones Significativas 19
2.4.1 MORE 19
2.4.2 Resultados Previos Para la Expresión Génica 21
2.5 PaintOmics 22
2.5.1 Enriquecimiento por Asociación 23
2.5.2 Aplicación de PaintOmics a los Datos del YMC 24
2.6 Métodos Estadísticos 25
2.6.1 Análisis de Componentes Principales (PCA) 25
3. Resultados y Discusión 26
3.1 Expresión Diferencial 26
3.2 PaintOmics 27
3.2.1 Nuevo Enfoque de Enriquecimiento Funcional 27
3.2.2 Resultados del Enriquecimiento Funcional 29
3.3 Regulación del metabolismo: MORE 30
3.4 Regulación del metabolismo: PLS 33
3.5 Comparación modelos MORE y PLS 39
4. Conclusiones 41
5. Glosario 43
6. Bibliografía 44

Lista de Figuras

Figura 1. Fluctuación de los niveles de oxígeno (dO_2) a lo largo del YMC. Se muestran los 15 puntos temporales seleccionados para los análisis de RNA-Seq y ChIP-Seq. Las tres fases del YMC quedan definicas por un código de color: magenta para OX, verde para RB y azul para RC. La figura proviene de (Sánchez-Gaya, 2018). 9
Figura 2. Fluctuación de los niveles de oxígeno (dO_2) a lo largo del YMC. Se muestran los 24 puntos temporales seleccionados para los análisis de metabolómica para la técnica LC-MS/MS (izquierda) y GC-TOFMS (derecha). Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC. 10
Figura 3. Gráfica PCA. Se muestran los scores de los puntos temporales para los 50 metabolitos compartidos entre los protocolos de la técnica LC-MS/MS. El código de color determina el protocolo siendo; rosa: 0.1% de ácido fórmico (protocolo 1); verde: 5mM de NH4OAc (protocolo 2). El número indica el protocolo (1: protocolo 1; 2: protocolo 2) y el punto temporal (del 1 al 12). 12
Fgura 4. Diagrama de Cajas para los 12 puntos temporales de la técnica LC-MS/MS. En (A) se presentan los datos burtos; (B) los mismos datos tras la transformación logarítmica. Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC. 13
Figura 5. Gráfica PCA loadings. Se muestra la distribución de los loadings para cada punto temporal en la técnica LC-MS/MS, siendo: (A) datos brutos; (B) datos transformados logarítmicamente. 13
Figura 6. Gráfica PCA. Se muestran los scores de los puntos temporales para los metabolitos compartidos entre las técnicas LC-MS/MS y GCxGC-TOFMS. En rosa se representan los pertenecientes a LC-MS/MS (LC) y en verde GCxGC-TOFMS (GC). 13
Figura 7. Distribución de los datos faltantes (NA) en los datos de metabolómica (LC-MS/MS y GCxGC-TOFMS) tras el alineamiento con la referencia. En rojo se representan los datos faltantes y en azul los datos conocidos. A la derecha se define la propoción de datos conocidos/faltantes en la muestra. 15
Figura 8. Gráfica PCA. Se muestran los scores de los puntos temporales para la técnica LC-MS/MS antes de la imputación de valores faltantes (B) y tras la aplicación de MICE (A). En la A se señalan los puntos imputados (4,7,8 y 9). Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC. 16
Figura 9. Perfil temporal de la isoleucina detectada mediante la técnica GCxGC-TOFMS. A la izquierda se muestra el perfil antes de la imputación de valores faltantes por MICE. A la derecha tras la imputación. En rojo se representan los puntos temporales imputados y en verde los conocidos. 16
Figura 10. Perfil temporal de GSH detectada mediante la técnica LC-MS/MS. Se muestra el perfil original antes de la interpolación con imputeTS (rojo) y tras la misma. En verde se representa la interpolación mediante el método lineal, en amarillo spline y en azul con Stine. 17
Figura 11. Gráfica PCA. Se muestran los scores de los puntos temporales para la técnica LC-MS/MS tras la interpolación de datos faltantes mediante el protocolo Stine del paquete imputeTS. Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC. 18
Figura 12. Perfil temporal del fumarato detectado mediante la técnica GCxGC-TOFMS. La curva negra con puntos rojos representan los puntos temporales obtenidos tras la aplicación del protocolo stine. En verde, se muestra el mismo perfil tras la adición de ruido. 18

Figura 13. Diagrama de barras con el número total de genes asociados significativamente a cada una de las variaciones de histonas (eje Y). Cada uno de los colores hace referencia a una modificación de histona concreta. La figura proviene de (Sánchez-Gaya, 2018). 21
Figura 14. Captura de pantalla de la interfaz de PaintOmics. Se muestran los diferentes campos que pueden ser completados para una ómica regulatoria. 23
Figura 15. Tabla de contingencia para el enriquecimiento por asociación de PaintOmics. Cada columna hace referencia a las regulaciones significativas o no significativas. Cada fila a aquellas en las que interviene un gen que ha sido encontrado, o no, en una ruta KEGG concreta. Así, en este ejemplo, se estudian un total de 70197 regulaciones potenciales gen-TF. También se incluye el p-valor obtenido al aplicar un test exacto de Fisher a esta tabla de contingencia. 24
Figura 16. Clusters resultantes de la aplicación de maSigPro. Se representa el perfil medio de los metabolitos incluidos en cada uno de los clusters. 26
Figura 17. A la izquierda diagrama de barras con el número óptimo de componentes por modelo válido (47). A la derecha, histograma con la distribución del R^2 estos modelos. Estos datos corresponden a los modelos de MORE. 31
Figura 18. Patrón de Expresión de RAX2 (YLR084C) junto con el perfil del metabolito IMP (arriba) o uracilp (bajo) a lo largo del tiempo. 32
Figura 19. Patrón de Expresión de BDF1 (YLR399C) junto con el perfil del Isocitrato. Abajo se muestra el patrón de expresión de IQGL1 junto al mismo metabolito anterior. 33
Figura 20. A la izquierda diagrama de barras con el número óptimo de componentes por modelo. A la derecha, histograma con la distribución del R^2 de los modelos. Estos datos corresponden a los modelos antes de la selección de variables. 34
Figura 21. R^2 acumulado en función del número de variables por cada componente (Comp.1 y Comp.2) para el metabolito Acetil-CoA. Cada recta queda marcada por el número de componentes que genera el mayor valor de R^2 (estrella) y el número óptimo de componentes siguiendo el criterio SIMCA-P (rombo). 34
Figura 22. A la izquierda diagrama de barras con el número óptimo de reguladores por modelo. A la derecha, histograma con la distribución del R^2 de los modelos. Estos datos corresponden a los modelos PLS-sparse. 35
Figura 23. Se comparan los estadísticos R^2, MSEP y Q^2 entre los modelos PLS originales y su variante sparse. La línea discontinua delimita la mejora/empeoramiento de dicho estadístico cuando se produce la selección de variables. En el caso concreto de R^2, las lineas rojas señalan el umbral de 0.7 seleccionado para marcar la "validez" del modelo. 36
Figura 24. Histograma de la distribución de la correlación entre metabolito y gen. Los valores de la correlación están en valor absoluto. Estos datos corresponden al modelo Sparse-PLS. 37
Figura 25. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito IMP. A la derecha el perfil de los genes asociados al mismo por el modelo Sparse-PLS tras el filtro de correlación. 38
Figura 26. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito isocitrato. A la derecha el perfil de los genes asociados al mismo por el modelo Sparse-PLS tras el filtro de correlación. 38
Figura 27. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito carnitina. A la derecha el perfil de los genes asociados al mismo por el modelo Sparse-PLS tras el filtro de correlación. 39
Figura 28. Diagrama de cajas de la correlación entre los modelos y sus metabolitos. Los modelos comparados son: MORE (azul) y PLS-sparse (antes del filtro por correlación, rojo). Se compara la correlación en mínimo (min), máximo (max), media (mean) y mediana (med). 40

Lista de Tablas

Tabla 1. Gráfica Gantt con las tareas realizadas a lo largo de la PEC0. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *4*
Tabla 2. Gráfica Gantt con las tareas realizadas a lo largo de la PEC1. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *4*
Tabla 3. Gráfica Gantt con las tareas realizadas a lo largo de la PEC2. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *5*
Tabla 4. Gráfica Gantt con las tareas realizadas a lo largo de la PEC3. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *6*
Tabla 5. Gráfica Gantt con las tareas realizadas a lo largo de la PEC4. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *6*
Tabla 6. Gráfica Gantt con las tareas realizadas a lo largo de la PEC5. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas. *7*
Tabla 7. Resumen de los metabolitos totales encontrados para las técnicas LC-MS/MS, GCxGC-TOFMS así como los protocolos de extracción utilizados en la primera (Protocolo 1 y Protocolo 2). Se incluye el número de dichos metabolitos que han sido encontrados en ambas técnicas o protocolos. *11*
*Tabla 8. Tabla del alineamiento de los puntos temporales de cada técnica metabolómica (LC-MS/MS y GCxGC-TOFMS) a la referencia. Como referencia se toman las observaciones del estudio de Sánchez-Gaya (Sánchez-Gaya, et al., 2018) representadas en la Figura 1. NA: casos en los que el alineamiento con la referencia no fue posible; * situaciones en las que el punto de referencia se encuentra entre dos puntos de la técnica metabolómica, se calculó la media.* *14*
Tabla 9. Tabla con el p-valor ajustado y odds ratio para los trece factores de transcripción que regulan a un mayor número de genes. La tabla proviene de (Sánchez-Gaya, 2018). *21*
Tabla 10. Tabla de contingencia resumen de PaintOmics para los tres tipos de enriquecimiento disponibles: por gen, feature y asociación. Como relevantes se toman aquellos TFs seleccionados como significativos. *28*
Tabla 11. Rutas significativas obtenidas tras la aplicación del análisis de enriquecimiento en PaintOmics junto con el p-valor combinado. En este caso, dichas rutas están ordenadas ascendentemente en función de dicho p-valor. *29*
Tabla 12. Rutas más significativas obtenidas tras la aplicación del análisis de enriquecimiento en PaintOmics junto con el p-valor combinado y el p-valor del enriquecimiento para metabolómica. En este caso, dichas rutas están ordenadas ascendentemente en función del p-valor del enriquecimiento por metabolito. *30*

1. Introducción

1.1 Contexto y justificación del Trabajo

En los últimos años la biología de sistemas se ha convertido en una de las ramas más interesantes dentro de la investigación biológica. Y es que entender el funcionamiento celular como una red interactiva ha permitido descubrir nuevas propiedades emergentes que hasta el momento nos eran desconocidas (Horgan, 2011). Este planteamiento ha sido posible gracias al avance en las tecnologías ómicas como RNA-seq, ChIP-Seq o la cromatografía.

Sin embargo, pese a la creciente accesibilidad de dichas tecnologías, la comunidad científica aún se halla lejos de alcanzar el ideal prometido por la biología de sistemas. Para llegar a él, uno de los mayores retos está en el desarrollo de herramientas y protocolos estadísticos que permitan la integración de los datos generados por las mismas y faciliten su comprensión.

En este contexto, uno de los sistemas biológicos más estudiado es el ciclo metabólico de la levadura (YMC, del inglés *Yeast Metabolic Cycle*): un proceso respiratorio que aparece cuando la levadura es cultivada en un medio con nutrientes limitados y cuya duración oscila entre 4 a 5 horas (Kuang, 2017). Lo más interesante de dicho ciclo es la oscilación periódica en la expresión génica. La regulación detrás de esta aún es desconocida, aunque, basándonos en los últimos estudios, parece que podría ser consecuencia de la interacción entre las variaciones metabólicas y la estructura de la cromatina (Tu, 2005). Esto último convierte al YMC en un sistema perfecto para el estudio de la relación entre metabolómica, transcriptómica y cromatina desde un punto de vista integrativo.

En un trabajo previo, Sánchez-Gayá y colaboradores (Sánchez-Gayá, 2018) aprovechando las cualidades del modelo, desarrollaron un protocolo de integración estadística que permitía relacionar datos del estado de la cromatina (ChIP-Seq) y de la expresión de factores de transcripción con la expresión génica (RNA-Seq). Mediante el mismo, fueron capaces de encontrar nuevas interacciones nunca antes descritas entre ambas y señalaron al metabolismo como la siguiente pieza a añadir al modelo.

Y es que, como ya es bien conocido, cualquier proceso fisiológico permanece incompleto sin el conocimiento del metaboloma. Esto se debe a la gran conectividad que existe entre las redes metabólicas y su dependencia inherente entre la regulación enzimática, concentración de metabolitos y flujos (Nielsen, 2003).

Este trabajo fin de máster intentará solventar este problema mediante el desarrollo de un protocolo para la integración estadística de la metabolómica, transcriptómica y ChIP-Seq en el contexto del YMC. Con este modelo pretendemos corroborar los resultados expuestos en el trabajo de Sánchez-Gayá así como generar una metodología robusta de integración que pueda ser utilizada en los futuros datos metabolómicos que se obtendrán próximamente en el laboratorio de Genómica de la Expresión Génica del Centro de Investigación Príncipe Felipe, donde se ha desarrollado el presente trabajo.

Para conseguirlo, se utilizarán los datos de ChIP-Seq y RNA-Seq del trabajo de Sánchez-Gayá. Por otro lado, y a la espera de recibir los datos de metabolómica producidos por el grupo, se buscaron datos públicos de metabolómica sobre el YMC que pudieran ser integrables con los demás datos ómicos disponibles. Los datos seleccionados fueron los obtenidos en la

investigación de Mohler (Mohler, 2008), la cual recoge información metabolómica en dos YMC consecutivos empleando dos tipos de tecnología: LC-MS/MS y GCxGC-TOFMS.

Finalmente, de acuerdo con Kuang y colaboradores, consideraremos que el YMC puede estar separado en tres frases dependiendo del perfil de expresión. Concretamente: la oxidativa (OX), la reductiva/creación (RB) y la reductiva/cambio (RC). Es por ello que a lo largo del trabajo se hará mención a estas tres fases aceptando la tesis de Kuang (Kuang, 2014).

1.2 Objetivos del Trabajo

Los objetivos de este proyecto son básicamente dos:

1. Obtención y procesado de datos de metabolómica que complementen los datos ómicos previos del YMC.
2. Desarrollo de un protocolo para la integración estadística de datos de metabolómica, transcriptómica y estado de la cromatina para estudiar la relación entre estas capas de regulación en el contexto del YMC.

Los objetivos anteriores quedarán superados si se cumplen los siguientes objetivos secundarios:

En el caso del primer objetivo estos son:

- Obtención de datos metabolómicos a partir de la literatura o mediante simulación.
- Exploración de los datos metabolómicos mediante técnicas descriptivas.
- Tratamiento de los datos metabolómicos (de ser necesario), mediante transformaciones, normalización, escalado, etc.
- Adaptación de los datos de metabolómica al diseño experimental del resto de ómicas disponibles.
- Selección de los metabolitos que cambian a lo largo de las fases del YMC mediante métodos estadísticos apropiados.

En el caso de que el objetivo uno quede superado se procederá con el segundo objetivo. Los objetivos secundarios relacionados con este son:

- Selección de un modelo adecuado para la integración de datos ómicos. Se probarán métodos como modelos de regresión múltiple y PLS.
- Construcción de una red que resuma las interacciones entre las ómicas. De esta forma podrán ser analizadas, descritas y comprendidas.
- Interpretación biológica de los resultados. Así como su comparación con la información obtenida en estudios previos.

1.3 Enfoque y método seguido

Para completar los objetivos anteriores se discutieron diversas estrategias y aproximaciones. Principalmente los puntos clave analizados fueron la obtención de los datos metabolómicos y la metodología para la integración de los tres tipos de datos ómicos.

El primer punto deriva de los resultados obtenidos por el laboratorio. Y es que tras la integración de los datos de expresión génica con la variación de histonas, el estado metabólico celular se planteó como la siguiente pieza clave en la comprensión del ciclo. Es por ello que el grupo decidió replicar las condiciones del experimento de Kuang para la obtención de datos metabolómicos. Sin embargo, debido a restricciones temporales, estos datos no estarían disponibles para la presente tesis pero uno de los objetivos de este trabajo era la preparación de la estrategia estadística de integración que pudiera ser aplicada a los futuros datos. Para el desarrollo de dicha estrategia se requerían datos apropiados de metabolómica, por lo que se plantearon dos estrategias diferentes para obtenerlos: búsqueda de datos públicos o simulación de datos.

Entre las otras dos opciones se dio preferencia a la búsqueda en bases de datos públicas y se dejó la simulación de datos como una última opción. Tras el análisis de varios trabajos finalmente se decidió hacer uso de la investigación publicada por Mohler (Mohler, 2008). Esta fue seleccionada por presentar unas condiciones experimentales bastante similares a la investigación original de Kuang.

El segundo punto fue la elección de la estrategia de integración. Actualmente se han descrito dos aproximaciones mayoritarias en el campo de la integración de datos ómicos: la integración conceptual y la estadística (Cavill, 2016).

La integración conceptual abraza la idea del análisis de cada ómica por separado y una interpretación conjunta de los resultados. Aunque más sencilla, esta aproximación presenta varias limitaciones en el descubrimiento de nuevas interacciones entre los datos (Cavill, 2016).

Por otra parte la integración estadística está basada en la aplicación de diversos modelos estadísticos, como por ejemplo modelos multivariantes (PCA o PLS) o modelos regresión, sobre los datos ómicos (Cavill, 2016). Este tipo de integración requiere una mayor potencia computacional y matemática. Pese a ello, ofrece mejores resultados que la anterior siendo considerada como la verdadera integración. El principal problema de esta estrategia es la dificultad que plantea siendo la razón de la ausencia de herramientas y protocolos disponibles. Es por ello que con el objetivo de contribuir al desarrollo de estas, y teniendo en cuenta su mayor potencia, se decidió hacer uso de esta aproximación por encima de la primera.

Debido a su capacidad en la reducción de dimensiones, los modelos multivariantes suelen ser los más utilizados en estudios de integración ómica y es que, a diferencia de otros, estos trabajan muy eficientemente en situaciones donde el número de variables explicativas supera al número de observaciones (Trygg y Wold, 2002). Por su lado, los modelos de regresión clásicos requieren de una reducción previa en el número de variables mediante distintas técnicas de selección.

En este trabajo se decidió hacer uso de ambos modelos: modelos multivariantes y modelos de regresión lineal y comparar los resultados obtenidos con cada uno de ellos para evaluar sus ventajas y limitaciones.

En cuanto a los modelos multivariantes se utilizó el PLS. Mientras que en el caso de los modelos de regresión se empleó un paquete de R desarrollado por el grupo de Genómica de la Expresión Génica: MORE.

1.4 Planificación del Trabajo

En este apartado se muestran las diferentes tareas que se desarrollaron a lo largo de las PECS. Cada apartado incluye un breve resumen de las tareas y problemas encontrados, así como un diagrama de Gantt con la organización temporal de las actividades.

PEC0 (del 19/09/2018 al 01/10/2018)

A lo largo de este periodo me reuní en varias ocasiones con la tutora. En estas reuniones se discutieron los objetivos del trabajo y las estrategias a seguir para alcanzarlos. Se generó el resumen inicial del trabajo y se seleccionó un primer título. Finalmente se inició la búsqueda de los datos de metabolómica así como el estudio de la simulación de datos.

Tabla 1. Gráfica Gantt con las tareas realizadas a lo largo de la PEC0. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1	Semana 2
Reuniones para discutir las estrategias y el tema del trabajo		
Estudio de simulación de datos		
Búsqueda de artículos sobre metabolómica en YMC		

PEC1 (del 02/10/2018 al 15/102018)

En este periodo conseguimos encontrar los datos de metabolómica y comenzamos con su análisis. Es en esta PEC donde se inició la adecuación de los datos. Del mismo modo empezamos a discutir sobre los diferentes modelos estadísticos de integración. En un primer momento nos centramos en los basados en redes Bayesianas.

Tabla 2. Gráfica Gantt con las tareas realizadas a lo largo de la PEC1. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1	Semana 2
Análisis descriptivo de los datos		
Transformación de los datos		
Estudio de estrategias de integración de datos ómicos		
Elaboración del documento del plan de trabajo		
Entrega del documento del plan de trabajo		

PEC2 (del 16/10/2018 al 19/11/2018)

Ante todo durante este tiempo intentamos resolver los problemas encontrados durante el procesamiento de los datos metabolómicos. Entre ellos cabe destacar los relacionados con valores faltantes. Y es que la naturaleza de los mismos no era aleatoria por lo que muchas herramientas analizadas no fueron útiles. Paralelamente a ello, proseguimos con el estudio de modelos estadísticos de integración basados en modelos de regresión lineal, como el MORE.

Tabla 3. Gráfica Gantt con las tareas realizadas a lo largo de la PEC2. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5
Estudio métodos de Imputación					
Imputación de datos					
Análisis y comparación de métodos de Imputación					
Estudio de modelos de regresión lineal (MORE y maSigPro)					
Estudio de métodos de interpolación					
Aplicación de la interpolación					

PEC3 (del 20/11/2018 al 17/12/2018)

A lo largo de estas semanas nos centramos en el análisis de los datos interpolados y en la finalización de la construcción de la matriz final de datos de metabolómica. Además, fue en este periodo donde comenzamos con el análisis final de los datos. Para ello aplicamos estrategias de enriquecimiento mediante PaintOmics, así como modelos estadísticos de integración de datos. Concretamente, se probaron modelos multivariantes como PLS-PM y PLS y O2-PLS, y modelos de regresión.

Cabe destacar que a lo largo de este periodo encontramos grandes problemas con uno de estos modelos, concretamente la aproximación del PLS-PM, que requiere partir de un modelo biológico teórico. Tras intentar generar esta red biológica a partir de las rutas de KEGG para su aplicación se decidió descartar dicha estrategia del trabajo final porque, entre otras dificultades, no fue posible generar una red completa debido a que muchas variables ómicas (especialmente metabolitos) no habían sido medidas en nuestros datos.

Finalmente, se aplicaron los demás modelos propuestos y se analizaron los resultados para extraer conclusiones. También utilizamos estos resultados para comparar el modelo de regresión clásico con el modelo multivariante.

Tabla 4. Gráfica Gantt con las tareas realizadas a lo largo de la PEC3. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1	Semana 2	Semana 3	Semana 4
Análisis de datos Interpolados				
Generación de Ruido				
Selección de metabolitos por maSigPro				
Estudio de PLS-PM, O2-PLS y PLS				
Estudio de Paintomics				
Adaptación de todos los datos a Paintomics				
Análisis resultados Paintomics				
Elaboración del modelo de red para PLS-PM				
Discusión de problemas encontrados				
Análisis e interpretación resultados de maSigPro				
Aplicación de MORE				
Aplicación de PLS				
Interpretación de los resultados de MORE				
Interpretación de los resultados de PLS				
Escribir la memoria del TFM				

PEC4 (del 18/12/2018 al 02/01/2019)

Finalmente en este intervalo de tiempo nos centramos en la escritura de la memoria y su corrección. Así como la finalización de los últimos detalles del trabajo y la preparación de la presentación oral.

Tabla 5. Gráfica Gantt con las tareas realizadas a lo largo de la PEC4. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1	Semana 2
Escribir memoria de TFM		
Análisis alternativos		
Preparación presentación oral del TFM		

PEC5 (del 03/01/2019 al 10/01/2019)

Para terminar con esta asignatura se procedió a elaboración de la presentación oral de tesis. Estas semanas fueron dedicadas exclusivamente a ello pues al haber entregado ya la memoria de prácticas, fue imposible proseguir con la investigación en el contexto de la asignatura.

Tabla 6. Gráfica Gantt con las tareas realizadas a lo largo de la PEC5. Las celdas azules representan la semana en la cual se desarrollaron las actividades expuestas.

Tareas	Semana 1
Preparación de la presentación oral del TFM	
Práctica de la presentación oral	
Defensa pública del TFM	

1.4.1 Riesgos y puntos clave en el trabajo

A lo largo de este apartado se muestran los puntos clave que se deben conseguir para que el trabajo se concluya de forma exitosa. Así mismo, también se presentan algunos posibles riesgos asociados a ellos y se discuten algunas estrategias de mitigación.

En primer lugar, uno de los puntos críticos más importantes en este trabajo es la obtención de los datos de metabolómica. Y es que estos datos son imprescindibles para la ejecución de la investigación. Para asegurar su disponibilidad, se tuvieron en cuenta dos estrategias. Como ya se comentó en apartados anteriores, se optó por la búsqueda de datos públicos. Sin embargo, en el caso de no haber encontrado ningún estudio adecuado, se habría procedido a la simulación de los mismos. De forma que, sea como sea, los datos habrían estado disponibles para la tesis de máster.

Por otro lado, otro punto es la adaptación de los datos al resto de ómicas. El riesgo de no conseguir este objetivo habría significado la imposibilidad de realizar una integración estadística y se tendría que haber optado por la integración conceptual.

Finalmente, el último punto crítico es la selección de un modelo de integración estadística para los datos. Los riesgos de este punto son los mayores ya que al tratarse de un modelo nuevo son muchos los problemas que pueden derivar en el transcurso del mismo. Una forma de solventar dichos problemas sería o bien cambiando de modelo, simplificando lo o presentar lo que se ha conseguido hasta el momento y proseguir con su estudio una vez finalizado el máster.

1.5 Breve sumario de productos obtenidos

A la finalización de esta tesis se esperan haber generado los siguientes documentos:

- Los archivos pertenecientes a las **PEC0**, **PEC1**, **PEC2** y **PEC3**. Los cuales incluirán la descripción del TFM (PEC0), el plan de trabajo (PEC1), y la descripción de las tareas realizadas a lo largo del mismo (PEC2 y PEC3).
- La memoria del TFM. Este documento incluirá toda la información del TFM incluyendo una introducción, métodos, resultados, conclusiones, etc.
- Un informe de autoevaluación.

Además de ello hay algunos documentos cuya generación dependerá de los resultados que se obtengan a lo largo del trabajo. Estos son:

- Un protocolo de integración de datos de metabolómica (LC-MS/MS, GCxGC-TOFMS), transcriptómica (RNA-Seq) y estado de la cromatina (ChIP-Seq). Este englobará todo el código de R utilizado. Es posible que dicho código sea posteriormente publicado en carpetas públicas.
- Una publicación científica que incluya parte del trabajo desarrollado.

1.6 Breve descripción de los otros capítulos de la memoria

De acuerdo con el plan docente el presente trabajo incluye los siguientes apartados:

- **Introducción**, en este apartado se justifica la necesidad del proyecto así como su contexto. En este caso la necesidad de desarrollar nuevos métodos para la integración estadística de datos ómicos y su aplicación en el estudio del YMC.

- **Materiales y Métodos**, a lo largo de esta parte se hará una descripción de los datos utilizados, el diseño experimental y la metodología utilizada en el trabajo.

- **Resultados y Discusión,** aquí se incluirán los resultados obtenidos en el apartado anterior y su interpretación basada en las evidencias científicas actuales.

- **Conclusión,** finalmente el apartado conclusión se dedicará a exponer los puntos más relevantes encontrados durante los resultados y discusión. Así como las dificultades encontradas a lo largo del trabajo.

2. Materiales y Métodos

2.1 Datos Ómicos

Los datos utilizados a lo largo de este proyecto son de tres tipos: datos metabolómicos adquiridos mediante las técnicas LC-MS/MS y GCxGC-TOFMs, datos de expresión génica obtenidos por RNA-Seq y, finalmente, datos de modificación de histonas mediante tecnología ChIP-Seq.

Los datos de metabolómica se descargaron desde la información suplementaria del artículo de Mohler (Mohler, 2008). En cuanto al resto, se obtuvieron del trabajo de Sánchez-Gaya y colaboradores (Sánchez-Gaya, 2018). Es importante apuntar que, aunque extensamente modificados por Sánchez-Gaya, estos últimos datos proceden a su vez del trabajo original de Kuang (Kuang, et al., 2014). Los datos originales de Kuang pueden ser encontrados en la base de datos GEO (Edgar, et al.,2002), con el código GSE52339.

2.1.1 Datos RNA-Seq

Los datos de RNA-Seq incluyen información de 5989 genes a lo largo de 15 puntos temporales en un YMC (Sánchez-Gaya, 2018)(**Figura 1**). Los datos de expresión fueron tomados mediante la plataforma HiSeq 2000 (Kuang, et al., 2014) y procesados, alineados, transformados y normalizados por Sánchez-Gaya. Esta normalización incluyó la estandarización por el número total de lecturas por muestra, la transformación logarítmica de los datos y su centrado.

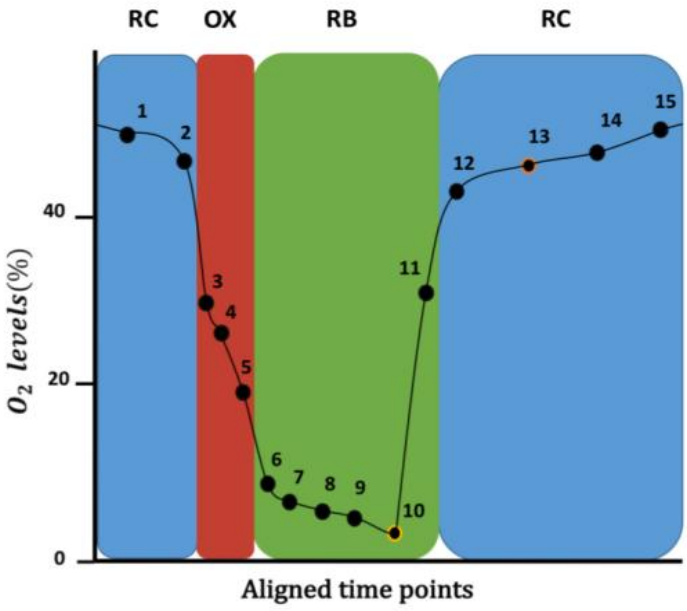


Figura 1. Fluctuación de los niveles de oxígeno (dO_2) a lo largo del YMC. Se muestran los 15 puntos temporales seleccionados para los análisis de RNA-Seq y ChIP-Seq. Las tres fases del YMC quedan definicas por un código de color: magenta para OX, verde para RB y azul para RC. La figura proviene de **(Sánchez-Gaya, 2018)**.

2.1.2 Datos ChIP-Seq

Del mismo modo, los datos de ChIP-Seq incluyen 6 variaciones de histonas a lo largo de 15 puntos temporales en un YMC (Sánchez-Gaya, 2018) (**Figura 1**). Estas modificaciones son: H3K9ac, H3K18ac, H3K4me3, H3K56ac, H3K14ac y H4K5ac. Para cada una de estas modificaciones Sánchez-Gaya obtuvo la cobertura por nucleótido a lo largo de todo el genoma. Tras ello, se definieron dos regiones genómicas por gen: La primera de 300 pares de bases (pb) desde el punto de inicio de la transcripción (por sus siglas en inglés, TSS) dirección *upstream*, y la segunda de la misma longitud pero dirección *downstream* desde el TSS. Finalmente, para cada una de estas dos regiones genómicas, Sánchez-Gaya calculó su cobertura media.

La definición de las regiones se realizó mediante el uso del archivo de anotación del genoma (de sus siglas en inglés, gtf) para *Saccharomyces cerevisiae*. Concretamente se descargó de Ensembl (Zerbino, et al., 2018) la versión 91.

Dado que en sus resultados Sánchez-Gaya encontró una elevada correlación entre los valores de ambas regiones, se decidió calcular la media de los valores de ambas. De esta forma, obtuvimos una única matriz de datos para cada modificación de histonas.

Finalmente, del mismo modo que en RNA-Seq, estos datos están normalizados, centrados y transformados logarítmicamente.

2.1.3 Datos Metabolómicos

A falta de los datos metabolómicos experimentales, este trabajo utilizó los datos publicados por Mohler (Mohler, 2008). En su investigación Mohler analizó el metaboloma a lo largo de dos YMC consecutivos. Para cada uno de estos ciclos se obtuvieron un total de 12 puntos temporales en los cuales se aplicaron dos técnicas de cromatografía distintas: LC-MS/MS (**Figura 2A**) y GCxGC-TOFMS (**Figura 2B**).

Adicionalmente, aunque en su trabajo Mohler únicamente hace mención a un protocolo de extracción de metabolitos mediante la técnica LC-MS/MS, los datos descargados incluían dos protocolos: (A) 0.1% de ácido fórmico y (B) 5mM de NH4OAc. En este trabajo se analizó la información de ambos protocolos y se realizó un análisis comparativo entre ambos.

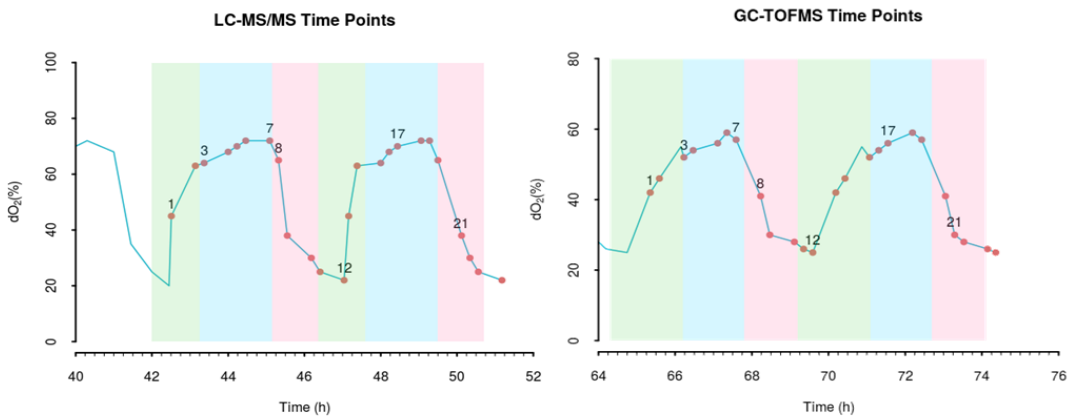


Figura 2. Fluctuación de los niveles de oxígeno (dO_2) a lo largo del YMC. Se muestran los 24 puntos temporales seleccionados para los análisis de metabolómica para la técnica LC-MS/MS (**izquierda**) y GC-TOFMS (**derecha**). Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC.

Finalmente, un resumen del número de metabolitos detectados por cada técnica y protocolo puede ser revisado en la **Tabla 7.**

Tabla 7. Resumen de los metabolitos totales encontrados para las técnicas LC-MS/MS, GCxGC-TOFMS así como los protocolos de extracción utilizados en la primera (Protocolo 1 y Protocolo 2). Se incluye el número de dichos metabolitos que han sido encontrados en ambas técnicas o protocolos.

	LC-MS/MS		GCxGC-TOFMS
	Protocolo 1	Protocolo 2	
Metabolitos Totales	65	50	45
Compartidos por Protocolo	50		-
Compartidos por Técnica	11		
Metabolitos Totales	110		

2.2 Preprocesado de los Datos de Metabolómica

Antes de proceder con la integración estadística de las ómicas, fue necesaria la aplicación de un protocolo de preprocesamiento sobre los datos de metabolómica. Este es un proceso esencial en cualquier tipo de estudio y recoge, principalmente, un análisis exploratorio de los mismos para deducir el tipo de transformaciones, normalizaciones, escalado, etc, necesarios para el correcto uso de ellos.

Una vez desarrollado el preprocesado se inició un protocolo de adaptación de los datos a las condiciones experimentales del resto de ómicas. Y es que, como se comentó al principio, las diferencias experimentales entre estudios deben ser resueltas antes de su integración. Entre estas diferencias destacan el número de YMC estudiados y la diferencia en los puntos temporales entre estudios. A lo largo de este apartado se abordan ambos pasos, el procesado y la adaptación de los datos para su integración.

2.2.1 Análisis Exploratorio

Con el objetivo de adaptar el diseño experimental de los datos de metabolómica al del resto de ómicas, diversos cambios fueron aplicados a los datos originales de Mohler. Uno de estos cambios consistió en la reducción del número de YMC, ya que había dos en el caso de metabolómica y uno para el resto de ómicas. Para ello, se tomó como referencia la **Figura 2** y se emparejó cada punto del primer ciclo con su correspondiente del segundo. A partir de ellos se calculó la media entre parejas reduciendo de 24 a 12 los puntos temporales.

Tras la reducción de los puntos de tiempo, se procedió a la selección de uno de los protocolos en la técnica LC-MS/MS. Y es que, como se describió en el apartado correspondiente, Mohler utilizó dos tipos de protocolos de aislamiento de metabolitos. Para decidir qué protocolo era mejor utilizar, se hicieron varios análisis.

En primer lugar, se aplicó un análisis de correlación entre los metabolitos que se habían

medido en ambos protocolos (**Tabla 7**). Los resultados mostraron una alta correlación entre ellos donde el 54% de los mismos presentaban una correlación del 90% mientras que únicamente en un 10% se observó una correlación menor del 50%. Para corroborar estos resultados, se realizó un análisis de componentes principales (PCA, del inglés *Principal Component Analysis*). Este PCA emparejó a la perfección cada una de las observaciones temporales independientemente de su procedencia (**Figura 3**) reafirmando los resultados anteriores. Dada esta alta correlación, la decisión fue tomada en base al número de metabolitos medidos en cada protocolo (65 frente a 50, **Tabla 7**), seleccionando al protocolo con 0.1% de ácido fórmico.

Una vez seleccionado el protocolo para la técnica LC-MS/MS, y reducido el número de ciclos, se procedió al análisis de la distribución de los datos mediante un gráfico de diagrama de cajas. Las distribuciones observadas (**Figura 4A**) presentaban una clara asimetría y varios valores atípicos. Para su corrección, se aplicó una transformación logarítmica sobre los datos. La distribución tras la transformación puede ser analizada en la **Figura 4B**.

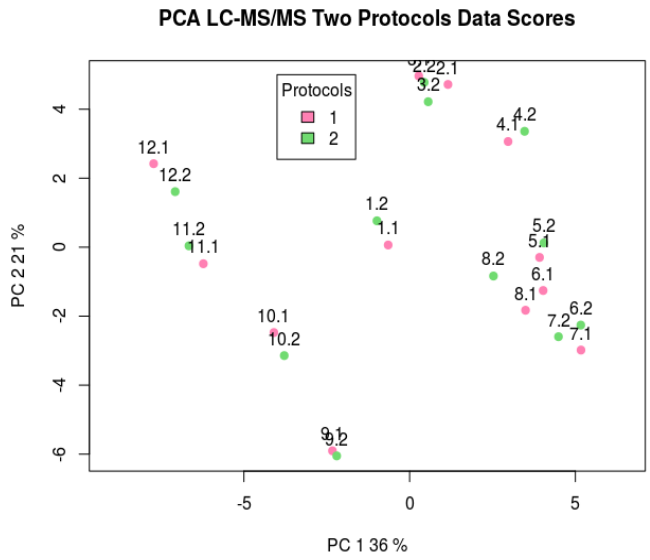
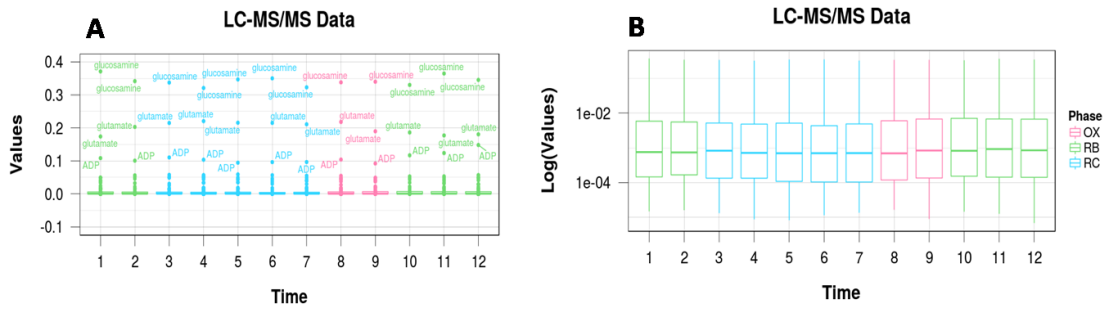


Figura 3. Gráfica PCA. Se muestran los *scores* de los puntos temporales para los 50 metabolitos compartidos entre los protocolos de la técnica LC-MS/MS. El código de color determina el protocolo siendo; rosa: 0.1% de ácido fórmico (protocolo 1); verde: 5mM de NH4OAc (protocolo 2). El número indica el protocolo (1: protocolo 1; 2: protocolo 2) y el punto temporal (del 1 al 12).

Adicionalmente, se aplicó un PCA sobre los datos transformados y originales. La mayor homogeneidad en la distribución de los loadings tras la transformación (**Figura 5**) apoyó el uso de esta.

Finalmente se analizaron los metabolitos compartidos entre técnicas mediante PCA. Para ello, previamente se centró la media a cero y se escaló cada variable para presentar una varianza igual a uno. Los resultados observados indicaron una ausencia de emparejamiento entre observaciones (**Figura 6**). Esta podría deberse a las diferencias entre los puntos temporales para cada una de las técnicas (**Figura 2**) o por incoherencias en los datos. Con tal de asegurar la primera opción, se decidió mantener todos los 11 metabolitos compartidos en la matriz final de datos incluyendo, para cada, una etiqueta de referencia (LC o GC). Consecuentemente, la matriz final contiene un total de 99 metabolitos únicos de los cuales 11 se encuentran duplicados.



Fgura 4. Diagrama de Cajas para los 12 puntos temporales de la técnica LC-MS/MS. En (**A**) se presentan los datos burtos; (**B**) los mismos datos tras la transformación logarítmica. Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC.

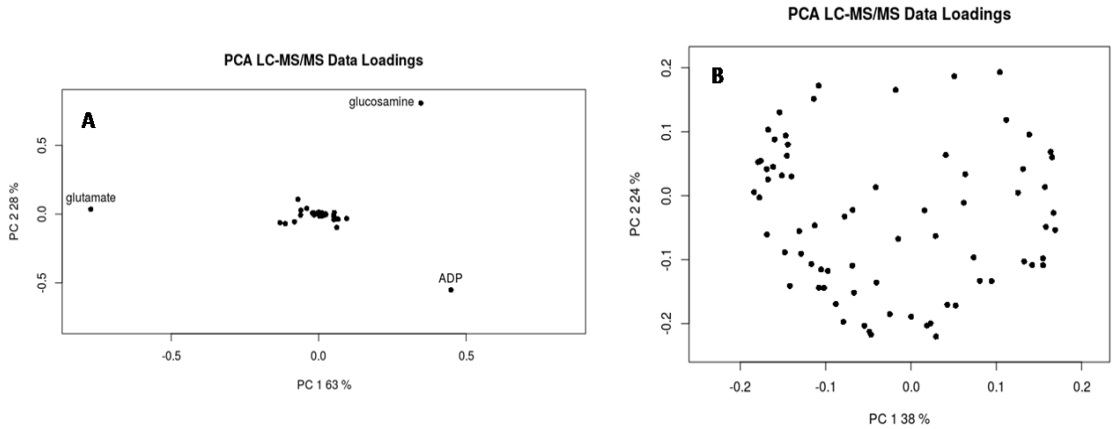


Figura 5. Gráfica PCA loadings. Se muestra la distribución de los loadings para cada punto temporal en la técnica LC-MS/MS, siendo: (**A)** datos brutos; (**B**) datos transformados logarítmicamente.

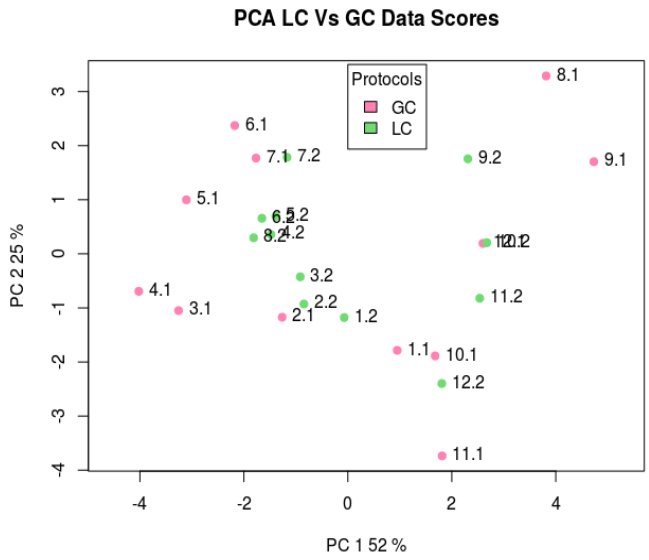


Figura 6. Gráfica PCA. Se muestran los scores de los puntos temporales para los metabolitos compartidos entre las técnicas LC-MS/MS y GCxGC-TOFMS. En rosa se representan los pertenecientes a LC-MS/MS (LC) y en verde GCxGC-TOFMS (GC).

2.2.2 Construcción de la Matriz Final de Datos de Metabolómica

Una vez los datos estuvieron preparados se inició la adaptación del diseño experimental de estos al de los estudios de ChIP-Seq y RNA-Seq, para que los puntos temporales fueran los mismos para todas las ómicas. Para ello, se tomaron como referencia las gráficas de RNA-Seq y ChIP-Seq (**Figura 1**) y se intentó realizar un alineamiento con los puntos de tiempo de metabolómica (**Figura 2**). Como resultado se obtuvo la **Tabla 8** en la cual se puede ver una gran cantidad de valores faltantes (NA) Estos valores fueron generados en aquellos casos en los cuales no fue posible emparejar las observaciones entre estudios.

Tabla 8. Tabla del alineamiento de los puntos temporales de cada técnica metabolómica (LC-MS/MS y GCxGC-TOFMS) a la referencia. Como referencia se toman las observaciones del estudio de Sánchez-Gaya (Sánchez-Gaya, et al., 2018) representadas en la **Figura 1**. NA: casos en los que el alineamiento con la referencia no fue posible; * situaciones en las que el punto de referencia se encuentra entre dos puntos de la técnica metabolómica, se calculó la media.

Referencia	LC-MS/MS	GCxGC-TOFMS
1	(6+7)/2*	7
2	8	NA
3	9	8
4	NA	9
5	10	10
6	11	NA
7	NA	NA
8	NA	11
9	NA	NA
10	12	12
11	1	(1+2)/2*
12	2	3
13	3	4
14	4	5
15	5	6

2.2.3 Datos Faltantes

Dada la importancia de mantener el máximo número de puntos temporales posible para la integración, se optó por el tratamiento de los datos faltantes en aquellos puntos de tiempo para los que no se disponía de medidas de metabolómica. Dos técnicas fueron testadas para la imputación de valores faltantes: estimación de valores faltantes e interpolación de datos.

La estimación de valores faltantes utiliza la información de los valores conocidos para la generación de modelos que predicen los valores faltantes (codificados como NAs). Dentro de esta metodología existen diversas aproximaciones, entre ellas las recogidas en el paquete MICE (Multivariate imputation by chained equations) y Amelia (Honaker, 2011) de R.

El algoritmo de Amelia está especializado en series temporales. La base matemática combina la toma aleatoria de muestras mediante "bootstrap" y el algoritmo EM ("expectation-maximization") sobre las mismas.

La combinación de ambas estrategias es básica para la función de Amelia. Y es que el algoritmo EM por si solo es incapaz de predecir los estimadores de máxima verosimilitud en situaciones de imputación múltiple. Sin embargo, el bootstrap resuelve el problema tomando m muestras de los datos originales y realizando la estimación de la media y la varianza. Tras ello, estas estimaciones son utilizadas en un análisis de regresión que genera los valores faltantes.

Si bien este algoritmo parecía encajar a la perfección con el problema; la naturaleza no aleatoria de los NAs (**Figura 7**), junto al hecho de que el método no admitía mayor número de variables que de observaciones (como es nuestro caso), impidió su aplicación. Como resultado se pasó a probar un modelo distinto de imputación. Concretamente el ofrecido por el paquete MICE.

MICE incluye una aproximación distinta. En este caso se producen una serie de iteraciones, cada una de ellas genera un modelo de regresión múltiple tomando una de las variables como respuesta y el resto como explicativas. Por cada iteración el modelo predice los datos faltantes de la variable respuesta hasta completar el set.

La calidad de la imputación fue analizada por PCA y mediante el análisis de los perfiles de alguno de los metabolitos a lo largo del tiempo.

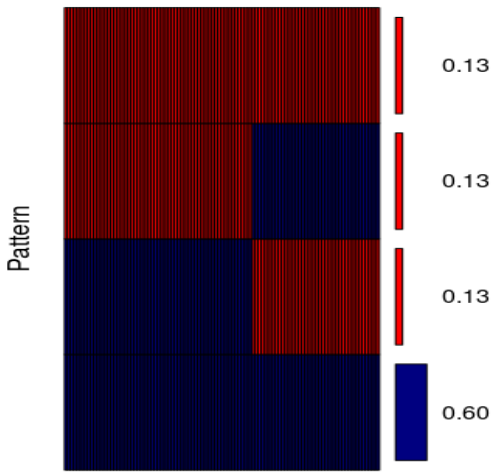


Figura 7. Distribución de los datos faltantes (NA) en los datos de metabolómica (LC-MS/MS y GCxGC-TOFMS) tras el alineamiento con la referencia. En rojo se representan los datos faltantes y en azul los datos conocidos. A la derecha se define la propoción de datos conocidos/faltantes en la muestra.

El primero de estos análisis mostró una mala distribución de los datos imputados. Se pudo ver que estos datos quedaban agrupados en el centro de la gráfica (**Figura 8**) sugiriendo grandes diferencias con las observaciones originales.

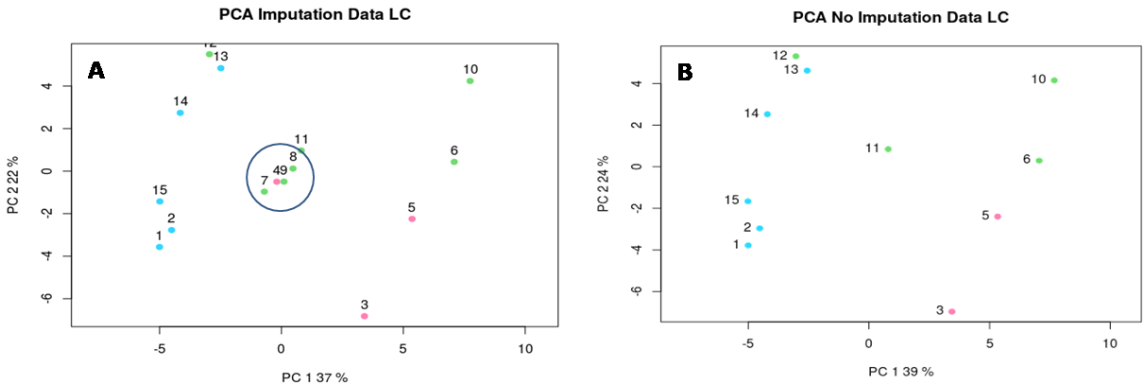


Figura 8. Gráfica PCA. Se muestran los scores de los puntos temporales para la técnica LC-MS/MS antes de la imputación de valores faltantes (**B**) y tras la aplicación de MICE (**A**). En la **A** se señalan los puntos imputados (4,7,8 y 9). Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC.

Del mismo modo, los perfiles de estos metabolitos se vieron significativamente alterados cuando los datos imputados eran tenidos en cuenta (**Figura 9**). Esto suponía un problema pues cambiaba directamente la interpretación del comportamiento del mismo en el YMC.

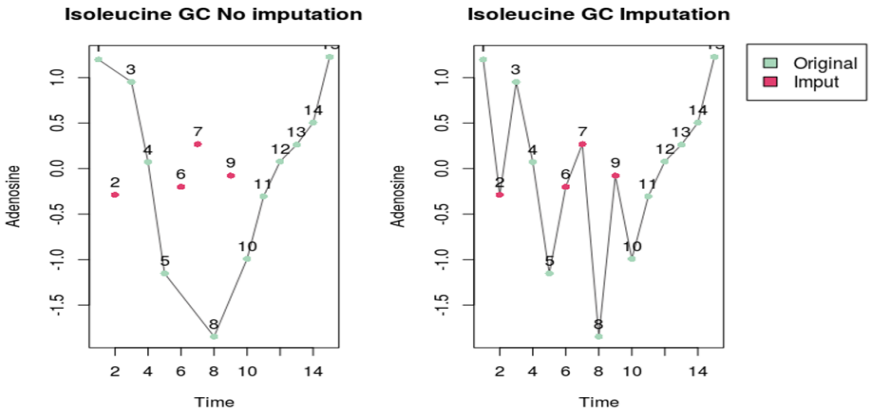


Figura 9. Perfil temporal de la isoleucina detectada mediante la técnica GCxGC-TOFMS. A la izquierda se muestra el perfil antes de la imputación de valores faltantes por MICE. A la derecha tras la imputación. En rojo se representan los puntos temporales imputados y en verde los conocidos.

En conjunto, ambos análisis demostraron que la estimación de valores faltantes era inadecuada para el objetivo propuesto. Por ello, se buscó una nueva estrategia, en este caso la interpolación de valores.

La interpolación utiliza una aproximación más sencilla a la resolución del problema: no genera un modelo lineal entre las variables sino que "dibuja" una recta a partir de los puntos conocidos. A partir de esta recta se estiman los valores faltantes. La generación de esta recta puede ser hecha de diversas formas, en este trabajo se testaron las incluídas en el paquete de R imputeTS (Moritz, 2017).

ImputeTS incluye varias propuestas para abordar el problema de valores faltantes. Entre ellas está la función *na.interpolation* que ofrece tres métodos de interpolación: lineal (*linear*), spline (*spline*) o "Cubic spline "(*stine*).

El primer método dibuja una recta lineal entre todos los puntos y a partir de esta interpola los valores de las observaciones desconocidas. El segundo y tercero utiliza curvas suavizadas basadas en modelos polinomiales. La diferencias entre estas dos últimas está en que la segunda asegura la preservación de la monotonicidad de la recta.

Para seleccionar la aproximación más eficiente, se testaron las tres estrategias sobre un grupo de metabolitos seleccionados al azar. En este caso se utilizó el análisis de los perfiles metabólicos para dicha comparación.

Lo que se observó es que los perfiles se veían mucho menos afectados que los vistos tras la aplicación de MICE (**Figura 10**). De forma concreta, el método lineal no variaba nada el perfil mientras que el método spline mantenía la tendencia pero con cambios en la pendiente bastante acusados. Finalmente, el método stine ofrecía un punto medio entre los otros dos. Por ese motivo, fue el seleccionado para este trabajo.

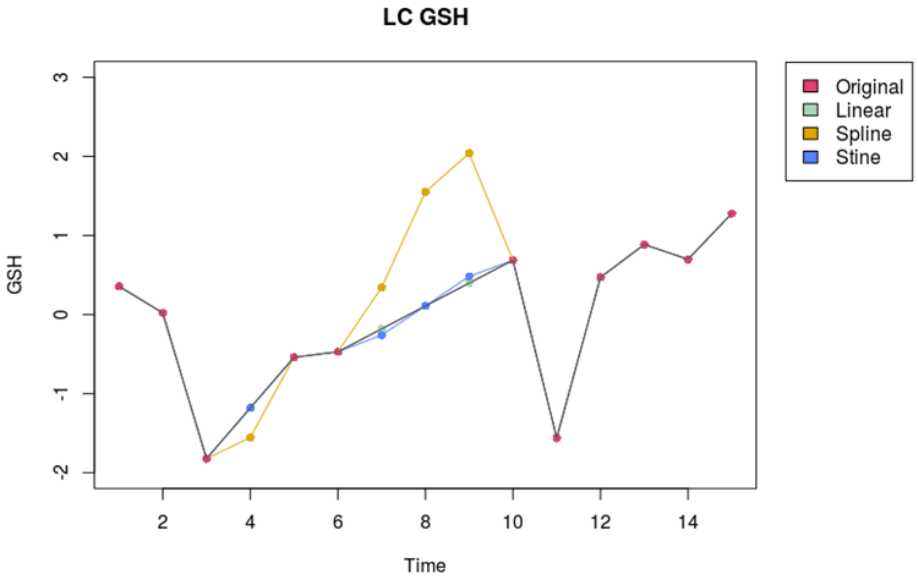


Figura 10. Perfil temporal de GSH detectada mediante la técnica LC-MS/MS. Se muestra el perfil original antes de la interpolación con imputeTS (rojo) y tras la misma. En verde se representa la interpolación mediante el método lineal, en amarillo *spline* y en azul con *Stine*.

Una vez seleccionado el método, se aplicó sobre la matriz de datos metabolómicos y se analizó mediante PCA. Lo que se observó fue que a diferencia del método de estimación, en este caso los puntos interpolados se repartían de forma correcta en el espacio (**Figura 11**).

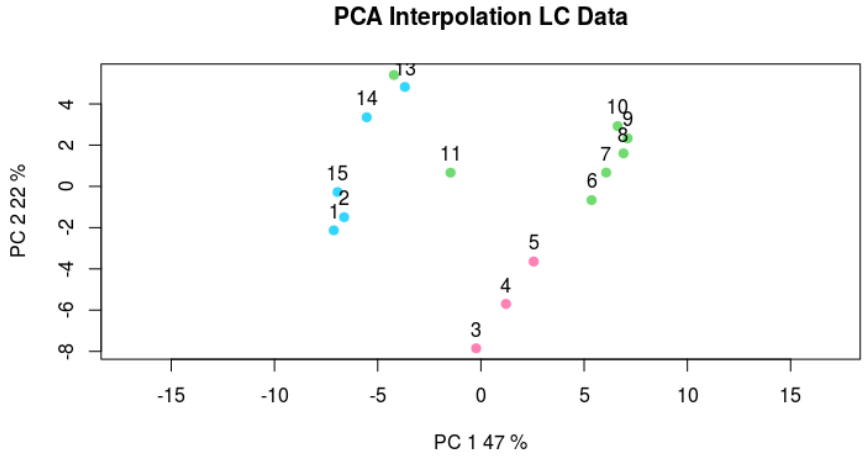


Figura 11. Gráfica PCA. Se muestran los scores de los puntos temporales para la técnica LC-MS/MS tras la interpolación de datos faltantes mediante el protocolo *Stine* del paquete imputeTS. Las tres fases del YMC quedan definidas por un código de color: magenta para OX, verde para RB y azul para RC.

Para que los datos interpolados fueran más realistas, se decidió añadirles ruido aleatorio. Dicho ruido se simuló a partir de una población normal con media cero y desviación típica igual a la diferencia entre los puntos adyacentes conocidos del valor interpolado dividida por dos. La ecuación utilizada pueden ser revisadas en la **Ecuación 1**.

$$N = (\mu = 0, sd = / \frac{b-a}{2} /) \qquad \textbf{Ecuación 1}$$

A partir de esta distribución se generaron aleatoriamente cinco valores y se calculó la media entre ellos, que se sumó al correspondiente dato interpolado. De nuevo, se estudiaron los perfiles de algunos metabolitos para verificar que no se hubiera producido ninguna anomalía durante este proceso (**Figura 12**). Tras comprobar con este análisis que todo había funcionado correctamente, el protocolo fue aplicado al grueso de los datos.

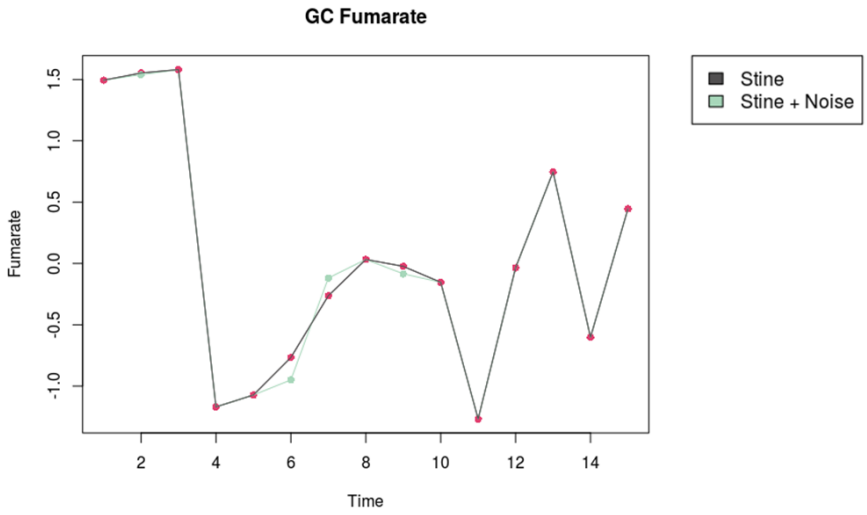


Figura 12. Perfil temporal del fumarato detectado mediante la técnica GCxGC-TOFMS. La curva negra con puntos rojos representan los puntos temporales obtenidos tras la aplicación del protocolo *stine*. En verde, se muestra el mismo perfil tras la adición de ruido.

2.3 Expresión Diferencial (maSigPro)

Para detectar la variación de algunos metabolitos a lo largo de los puntos de tiempo se aplicó la librería de R maSigPro (Conesa, 2006).

MaSigPro es una aproximación basada en modelos de regresión que permite el análisis de datos de series temporales (Conesa, et al., 2006). Aunque principalmente se pensó para el análisis de datos de microarray o RNA-seq, esta aplicación puede ser útil para la detección de cualquier variación cuantitativa a lo largo de una serie temporal. El objetivo de esta herramienta es la detección de variables ómicas (metabolitos en este caso) que varían significativamente a través del tiempo. Para ello se genera un modelo de regresión del perfil en función del tiempo y se analiza la existencia de variaciones significativas del mismo. Así mismo, en función de dicho perfil, el programa agrupa las variables (metabolitos) en un número determinado de *clusters* que fija el usuario. Estos *clusters* agruparán aquellas variables, genes o metabolitos, que presentan un perfil parecido.

Para obtener resultados comparables a los obtenidos para los datos de expresión génica, se decidió hacer uso de los mismos argumentos utilizados en el análisis de Sánchez-Gaya para la función de maSigPro. Se aplicó un nivel de significación de 0.05, el ajuste del p-valor se realizó mediante el procedimiento de Benjamini-Hochberg (Benjamini & Hochberg, 1995), el valor del coeficiente de determinación (R^2) mínimo exigido fue de 0.6 y el grado del polinomio exigido igual a 3.

Este procedimiento fue aplicado sobre los datos metabolómicos indicando la generación de un número total de tres *clusters*.

2.4 Regulaciones Significativas

El objetivo que se persigue en el presente estudio con la integración de datos ómicos es entender los mecanismos regulatorios del YMC. Así como en el trabajo de Sánchez-Gaya se analizó cómo los genes estaban regulados por factores de transcripción o modificaciones de histonas, en el presente trabajo se pretende estudiar la regulación de los metabolitos.

En ambos casos, la herramienta MORE nos permite alcanzar nuestro objetivo. En este apartado se describe en primer lugar el método MORE (cuyos resultados sobre los datos de metabolómica se mostrarán en la sección de Resultados) y los resultados obtenidos por Sánchez-Gaya tras la aplicación de MORE a los datos de expresión génica, ya que dichos resultados serán utilizados en la herramienta PaintOmics (ver sección 2.5).

2.4.1 MORE

El paquete de R MORE (del inglés, *Multi-Omics Regulation*) fue desarrollado por el laboratorio de Genómica de la Expresión Génica y está disponible en https://bitbucket.org/account/user/ConesaLab/projects/MORE. Esta aproximación hace uso de modelos de regresión para encontrar relaciones entre la expresión génica y sus posibles reguladores (por ejemplo TFs). Esta metodología es capaz de encontrar aquellos reguladores que afectan de forma significativa a la variable respuesta dentro de un grupo de candidatos. La variable respuesta suele ser la expresión génica, pero se admiten más tipos de datos como por

ejemplo: concentración de proteínas, niveles de transcritos, etc.

MORE aplica modelos lineales generalizados (GLM) (Nelder y Wedderburn, 1972) que, a diferencia de los modelos de regresión clásicos, utilizan el método de máxima verosimilitud (en inglés, *maximum likelihood*) para la estimación de los coeficientes de regresión. De esta forma, acepta variables respuesta con una amplia gama de distribuciones estadísticas, como: distribuciones normales (microarrays de expresión), Poisson y binomial negativa (lecturas de RNA-Seq), etc.

MORE cuenta con estrategias orientadas a resolver problemas de dimensión de los datos, en especial los casos en los que el número de variables supera el número de observaciones. Entre estas estrategias se encuentran: Filtro de variabilidad, filtro de multicolinealidad y métodos de selección de variables (modelos de penalización y *stepwise*).

Como se describe en (Faraway, 2005) la multicolinealidad aparece cuando algunos predictores son, aproximadamente, una combinación lineal de otros. Como consecuencia, los errores estándar de los coeficientes estimados pueden aumentar, provocando una gran imprecisión en la predicción, por lo que es muy importante resolver este problema. La estrategia de MORE para reducir el problema de multicolinealidad consiste en analizar las correlaciones entre variables explicativas y agregar en un mismo grupos aquellas que presentan una elevada correlación. Se elige al azar un representante de este grupo, que será el único que se incluya en el modelo inicial, por lo que se calculará un único coeficiente beta por cada grupo.

En cuanto a la selección de variables se puede optar por la penalización o el *stepwise*, o incluso aplicar ambos. El protocolo de penalización incluido en MORE es *Elastic Net shrinkage* (Zou & Hastie, 2005), una aproximación que combina la estrategia de *Ridge* (Hoerl y Kennard, 1970) y Lasso (Tibshirani, 1996). La segunda aproximación, s*tepwise* (Draper & Smith, 1998), incluye varias opciones como: *Forward, backward, two ways forward* y t*wo ways backward*.

Finalmente, MORE presenta otras funcionalidades destinadas a la comprensión de los resultados. Entre ellas destacan: resúmenes de resultados a nivel global o por gen, gráficas de relación entre genes y sus reguladores significativos, etc.

Los datos que requiere MORE para su aplicación son sencillos. Por una parte una matriz con las variables respuestas y las matrices de predictores medidas sobre las mismas muestras (puntos de tiempo en nuestro caso). Por otro lado, MORE necesita una matriz de relación entre cada una de las entidades de la variable respuesta (por ejemplo genes) y sus potenciales reguladores. Así mismo, también es posible incluir una covariable en el modelo que indique el grupo o condición experimental al que pertenece cada observación (por ejemplo para comparar dos tratamientos, dos grupos de sanos y enfermos, etc.).

Para nuestro estudio los datos utilizados fueron los obtenidos tras la aplicación de maSigPro. Concretamente aquellos relativos a los genes y metabolitos diferencialmente expresados en RNA-Seq y metabolómica respectivamente. Como variable respuesta se seleccionaron los datos metabolómicos mientras que como predictores fueron introducidos los datos de expresión génica. Para la relación entre predictor/respuesta se decidió utilizar todos los genes como posibles reguladores de todos los metabolitos. Esta es una aproximación novedosa, ya que MORE está diseñado para partir solo de los reguladores potenciales de cada variable respuesta que se pueden obtener experimentalmente o a partir de bases de datos. Dado que no disponíamos de datos de regulaciones potenciales entre genes y metabolitos, decidimos incluir en el modelo de cada metabolito todos los genes. Así pues, se aprovechó esta circunstancia para ver cómo funcionaba MORE en este escenario.

2.4.2 Resultados Previos Para la Expresión Génica

Anteriormente Sánchez-Gaya utilizó la metodología MORE para estudiar la regulación de la expresión génica por factores de transcripción (TFs) y diferentes modificaciones de histonas (H3K14ac, H3K4me3, H3K56ac, H4K5ac, H3K9ac y H3K18ac).

En su trabajo, Sánchez-Gaya encontró una elevada cantidad de asociaciones significativas entre estas modificaciones de histonas y los genes. En su mayoría, estas correlaciones eran positivas y apoyaban la teoría de (Berger, 2007) sobre el rol en la activación transcripcional de las mismas. Sus resultados encontraron a H3K9ac y H3K18ac como reguladoras del mayor número de genes con 780 en la primera y 815 en la segunda. Así mismo, la H3K56ac parecía variar la cantidad de genes regulados en función de la fase del YMA presentando una mayor cantidad durante la fase RB. Sus resultados a nivel global pueden ser consultados en la **Figura 13**.

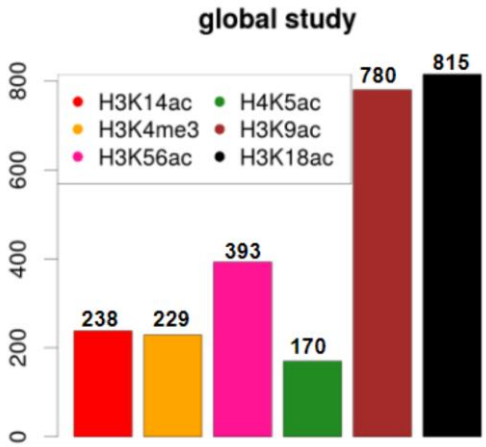


Figura 13. Diagrama de barras con el número total de genes asociados significativamente a cada una de las variaciones de histonas (eje Y). Cada uno de los colores hace referencia a una modificación de histona concreta. La figura proviene de **(Sánchez-Gaya, 2018)**.

Tabla 9. Tabla con el p-valor ajustado y *odds ratio* para los trece factores de transcripción que regulan a un mayor número de genes. La tabla proviene de **(Sánchez-Gaya, 2018)**.

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

En cuanto a los resultados correspondientes a los TFs, un total de 105 TFs fueron encontrados como reguladores significativos de 2480 genes. De estos, 13 TFs demostraron ser los que mayor proporción de genes regulaban (**Tabla 9**).

En conjunto, los resultados, tanto de asociaciones de histonas como los TFs fueron utilizados en el presente estudio asumiendo las decisiones tomadas y explicadas por Sánchez-Gaya en su trabajo (Sánchez-Gaya, 2018).

2.5 PaintOmics

La herramienta web PaintOmics v3.0 (Hernández-de-Diego, 2018) fue utilizada sobre los datos de transcriptómica, ChIP-Seq y metabolómica para estudiar el enriquecimiento funcional de las variables ómicas que resultaron significativas en los análisis anteriores (maSigPro y MORE).

PaintOmics es una herramienta web para la visualización integrativa de varios datos ómicos sobre rutas KEGG (*Kyoto Encyclopedia of Genes and Genomes*)(Dennis, et al., 2003). Las rutas KEGG representan abstracciones de rutas metabólicas concretas. En cada una de ellas se definen una serie de nodos que incluyen a aquellos genes que participan en alguna de las reacciones incluidas en la ruta. Así mismo, la interacción entre nodos queda definida por los substratos y productos de la reacción.

En este contexto, la aplicación acepta varios tipos distintos de datos: genes, metabolitos, proteínas, reguladores... En el caso de genes y proteínas el mapeo sobre las rutas es directo y únicamente es necesaria la transformación de los identificadores (IDs) a “Entrez” IDs, lo cual hace de forma automática la herramienta. Lo mismo ocurre con los metabolitos, en este caso PaintOmics busca dichos metabolitos sobre las distintas rutas KEGG. La problemática viene con los reguladores, ya que las rutas KEGG únicamente incluyen genes y metabolitos, y no reguladores, como TFs o variaciones de histonas. Para ello, la aplicación asocia cada regulador a los genes que potencialmente regula permitiendo así su mapeo sobre las rutas de KEGG.

Posteriormente, PaintOmics desarrolla diferentes análisis estadísticos de enriquecimiento funcional. Estos se aplican de forma individual a cada ómica y de forma combinada para todas ellas. En este último caso se pueden utilizar dos tipos distintos de test estadísticos, test de probabilidad combinada de Fisher o de Stouffer.

Para ello se requieren dos tipos de archivo por cada ómica. Por un lado la matriz de datos, por ejemplo la matriz de datos de expresión génica, y por otra aquellas entidades, siguiendo con el ejemplo genes, que son significativos, en este caso que varían su expresión de forma significativa. Así mismo, los valores contenidos en la matriz de datos deben representar una comparación (ratio) entre dos situaciones (*fold change)*, por ejemplo caso/control.

En cuanto al enriquecimiento individual por ómica, PaintOmics permite dos opciones distintas para el caso de los reguladores: por gen o por *regulador*. La diferencia entre ambas radica en la generación de la tabla de contingencia necesaria para aplicar el test exacto de Fisher que determinará si hay enriquecimiento de variables significativas en una determinada ruta. Mientras que el enriquecimiento por gen tiene en cuenta la aparición de cada gen, y cuenta cada gen solo una vez; la segunda cuenta el número de apariciones únicas del regulador. Para entenderlo mejor un ejemplo sería un factor de transcripción que afecta a varios genes distintos. En el caso de aplicarse el enriquecimiento por gen, se contarían todos aquellos genes

afectados por el TF, pero si hay más de un TF regulando a un gen, el gen solo se contaría una vez. Por otro lado, si se utilizase el enriquecimiento por regulador, un TF regulando a varios genes se contaría únicamente una vez pero si un gen está regulado por varios TFs , se contaría el número de TFs.

Tras sus cálculos PaintOmics devuelve una lista con las rutas donde las ómicas se encuentran enriquecidas. Además, esta lista viene acompañada del p-valor de cada ómica así como del p-valor combinado de todas ellas. A partir de esta lista, el usuario puede acceder a cada ruta y visualizar, de forma interactiva diferentes características. Concretamente, para cada ómica se puede observar: las *variables ómicas* mapeadas sobre la red, el cambio en ellas (representando en escala de rojo el aumento de la misma y en azul la disminución), y se destaca de forma especial la variable ómica en el caso de estar diferencialmente expresada.

2.5.1 Enriquecimiento por Asociación

Durante la aplicación de PaintOmics se encontraron algunas limitaciones relacionadas con las opciones de enriquecimiento para las ómicas reguladoras. Por ello, antes de su uso se procedió a la implementación de una nueva modalidad de enriquecimiento, por regulación.

Este enriquecimiento cambia la forma de calcular la tabla de contingencia, ya no se cuentan genes ni reguladores sino número de regulaciones o asociaciones gen-regulador. Para ello, se requiere disponer, no solo de las regulaciones potenciales, sino también de las regulaciones significativas. Estas regulaciones significativas se pueden obtener de distintas formas, por ejemplo, se pueden utilizar los resultados obtenidos por aplicaciones como MORE para determinar de aquellas interacciones descritas las que son significativas. Por ejemplo, en el caso de TFs, MORE determina cuáles de ellos son reguladores significativos de la expresión génica. En la **Figura 14** se puede ver la nueva interfaz de PaintOmics y cómo permite la selección de dicho enriquecimiento.

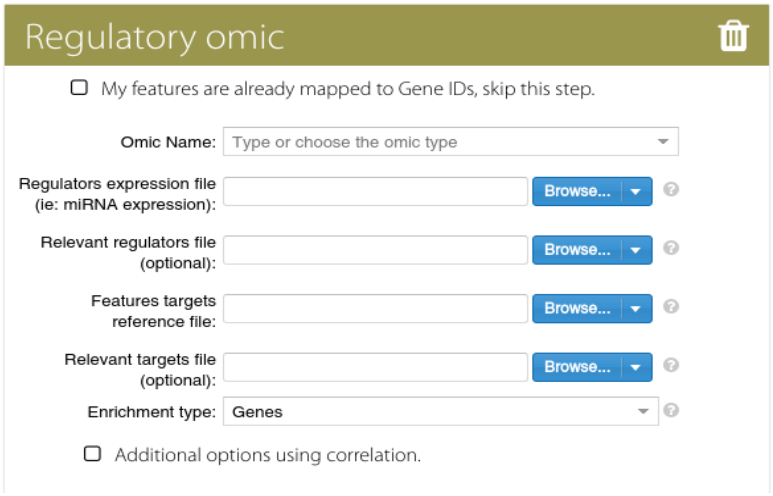


Figura 14. Captura de pantalla de la interfaz de PaintOmics. Se muestran los diferentes campos que pueden ser completados para una ómica regulatoria.

Finalmente, la **Figura 15** muestra un ejemplo de la tabla de contingencia resultante de la aplicación del enriquecimiento por asociación.

TF
p-value:0.01295

	Relevant	Not Relevant	
Found	3304	5189	8493
Not found	23228	38476	61704
	26532	43665	70197

Figura 15. Tabla de contingencia para el enriquecimiento por asociación de PaintOmics. Cada columna hace referencia a las regulaciones significativas o no significativas. Cada fila a aquellas en las que interviene un gen que ha sido encontrado, o no, en una ruta KEGG concreta. Así, en este ejemplo, se estudian un total de 70197 regulaciones potenciales gen-TF. También se incluye el p-valor obtenido al aplicar un test exacto de Fisher a esta tabla de contingencia.

2.5.2 Aplicación de PaintOmics a los Datos del YMC

Una vez implementada la nueva funcionalidad, se procedió a la aplicación de la herramienta. En este caso se utilizaron cuatro tipos datos: expresión génica, metabolómica, TFs y modificaciones de histonas (ChIP-Seq), tomando a las dos últimas como ómicas reguladoras.

Para todas ellas las matrices de *fold change* fue calculada tomando el punto de tiempo uno como referencia (**Ecuación 2**).

$$fold = log2(t) - log2(t1) \qquad \textbf{Ecuación 2}$$

Para la lista de genes/metabolitos significativos se utilizaron los resultados obtenidos por maSigPro. Y se seleccionó un tipo de enriquecimiento por gen en el caso de RNA-Seq y de *feature* para la metabolómica.

Por otra parte, el resto de datos se seleccionaron como ómicas reguladoras. En el caso de ChIP-Seq se incluyó cada variación como una ómica "reguladora" distinta. Para la lista de interacciones se supuso que dicha variación afectaba a todos los genes de RNA-Seq. Por otro lado la lista de interacciones significativas fue extraída de los resultados de MORE de Sánchez-Gaya.

Del mismo modo, en el caso de los TFs, la lista de interacciones se extrajo de la base de datos Yeastract (Teixeira, et al., 2018) mientras que las relaciones significativas de los resultados de MORE de Sánchez-Gaya.

El objetivo tras la aplicación de este análisis fue, por un lado, utilizar una herramienta adicional en el estudio de Sánchez-Gaya y observar si la adición de los datos metabolómicos alteraba significativamente sus conclusiones y, por otro lado, testar la nueva funcionalidad del programa.

2.6 Métodos Estadísticos

2.6.1 Análisis de Componentes Principales (PCA)

El PCA es un algoritmo matemático que permite la reducción del número de variables (dimensiones) en un estudio a partir de la creación de nuevas variables, componentes principales, que recogen la mayor parte de la variabilidad de la matriz original (Ringner, 2008). Cada una de estas componentes principales (PC) es combinación lineal de las variables originales y explica un porcentaje del total de variabilidad de los datos. Así mismo, cada PC está ordenada de acuerdo a dicho porcentaje de forma que la PC1 explica una mayor proporción que la PC2 y esta que la PC3. De esa forma se puede explicar la variabilidad de los datos con una selección reducida de dichas componentes.

Durante el cálculo de estas PC se generan los "loadings" y las "scores". Los primeros recogen la importancia que cada una de las variables originales tiene sobre la componente analizada. Por otro lado, los "scores" hacen referencia a la importancia de las observaciones en dichas componentes. La proyección de cada una de ellas en las nuevas componentes puede ser representada de forma gráfica. Esta gráfica contribuye a una mejor comprensión de su relación y puede servir para analizar si por ejemplo muestras del mismo grupo se emparejan en la misma zona del espacio o, si por el contrario, aparecen grupos de muestras no esperados.

A lo largo del trabajo este tipo de análisis se realizó principalmente para testar el comportamiento de los datos y asegurar que este no se veía comprometido por las variaciones aplicadas sobre los datos originales.

2.6.2 *Partial Least Square Regression* (PLS)

La regresión parcial de mínimos cuadrados (PLS) (Wold, et al., 2001), al contrario que el PCA, admite una matriz de datos como variable respuesta (por ejemplo metabolómica) y otra como predictora (como expresión génica). Así pues, el PLS es uno de los principales métodos para el análisis multivariante cuando se quiere comprobar la relación entre una matriz predictora (X) y una matriz respuesta (Y) (Trygg y Wold, 2002). Concretamente, es muy utilizado en el ámbito del estudio ómico. Esto se debe a la robustez del modelo frente a la colinealidad y al ruido, así como a su capacidad de análisis de bases de datos donde el número de variables supera al de observaciones (todo ello cualidades típicas de los datos ómicos).

Esta aproximación se basa en la reducción de dimensiones de los datos (Meng, et al., 2016). En un proceso similar al PCA, el PLS disminuye el número de variables observadas mediante la descomposición singular de cada una de las matrices de datos. A partir de esta descomposición se construyen las nuevas variables latentes. Durante este proceso el PLS busca encontrar la máxima covarianza entre el predictor y la respuesta (Kim-Anh, et al., 2008).

A partir del modelo generado es posible analizar la relación entre variables explicativas (reguladores en este caso) y respuesta (metabolitos) o utilizarlo para la predicción de la misma a partir de nuevos datos

En este caso, se utilizó el modelo de PLS implementado en el paquete de R MixOmics (Rohart, et al., 2017). Como variable respuesta se utilizaron cada uno de los metabolitos mientras que como explicativas los genes. Para cada metabolito se generó un primer modelo de PLS calculando el número óptimo de componentes. Para ello se generaron de forma automática

diferentes modelos con un número de componentes distinto, a partir de estos se analizó el R cuadrado en busca de aquel modelo que optimizara dicho parámetro. Para ello, se siguió el criterio SIMCA-P (Wold y Umetri, 1996). Este criterio puede ser aplicado tanto al R^2 como Q^2 acumulado durante la adición de nuevas componentes al modelo. El incremento de alguno de estos parámetros es calculado hasta que este no supere un 0.0975 ($R^2 > (1 - 0.95^2)$). Cuando se alcanza esta condición se supondrá que el modelo no mejora con el incremento de una componente extra.

Una vez conocido el número óptimo de componentes mediante el criterio anterior, se aplicó la variante *Sparse*-PLS.

Esta variante reduce el número de variables explicativas mejorando la interpretabilidad de los resultados en estudios biológicos. Este proceso se consigue mediante el desarrollo de una selección de variables utilizando la penalización LASSO (de las siglas en inglés, *least absolute shrinkage and selection operator*) en cada par de vectores *loadings* (Rohart, et al., 2017). De forma general, este método combina la contracción de algunos parámetros a cero y la selección de variables. Para ello se impone una penalización sobre los coeficientes de regresión calculados. En los casos que esta penalización alcanza cierto umbral se contrae el valor de β a 0 produciendo, consecuentemente, una selección de las variables explicativas más importantes.

Utilizando este método de penalización se seleccionaron las 10, 25, 50, 100 y 200 mejores variables por componente. A partir de los modelos resultantes, se calculó el R^2 acumulado. Una vez más, sobre este estadístico, se empleó el criterio SIMCA-P para seleccionar el modelo óptimo.

3. Resultados y Discusión

3.1 Expresión Diferencial

Se aplicó el método maSigPro para identificar los metabolitos que varían significativamente a lo largo del YMC. Se identificaron como diferencialmente expresados un total de 54 metabolitos de los 110 totales. Estos 54 metabolitos se agruparon en dos *clusters* distintos representados en la **Figura 16**.

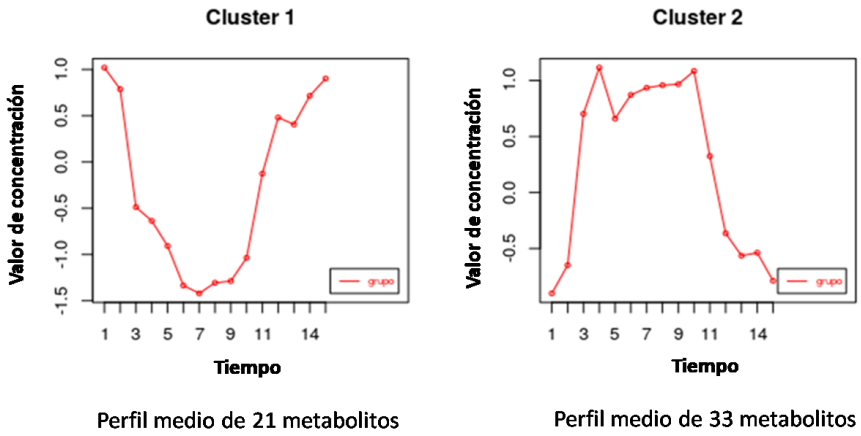


Figura 16. *Clusters* resultantes de la aplicación de maSigPro. Se representa el perfil medio de los metabolitos incluidos en cada uno de los clusters.

Como se puede observar, estos *clusters* parecen tener una relación directa con el YMC. Esto se puede comprobar comparando la **Figura 1** con la **Figura 16**. Y es que aparentemente la concentración de metabolitos se asocia directamente con la concentración de oxígeno del ciclo, y, consecuentemente, con una de las tres fases principales del mismo. De esta forma, en función del perfil medio de cada uno de ellos, es posible asociarlos a una fase concreta del ciclo. Por ejemplo, en el segundo *cluster* se puede ver un pico entre el tiempo 3 al 5, una pequeña subida a lo largo del tiempo 7 y 11 y finalmente un pico en el tiempo 12. Esto significa que dicho *cluster* representa dos fases distintas, la OX y la RB, mientras que hay un pequeño pico que marca el inicio de la fase RC. Por su parte, el primero de los *clusters* parece representar exclusivamente la fase RC.

En cuanto a los metabolitos asociados a cada *cluster* encontramos un 61% para el primero y un 39% para el segundo. Así mismo, la naturaleza de estos cuadra a la perfección con lo descrito en trabajos como (Tu, et al., 2007), (Tu, et al., 2005) o el propio Sánchez-Gaya.

Se observa como en el *cluster* dos se encuentran precursores de nucleótidos como GMP, Uracilo o IMP, intermediarios del ciclo del ácido carboxílico, como isocitrato y varios precursores para aminoácidos, por ejemplo homoserina o homocisteína. Estos aminoácidos han sido descritos como parte de la fase OX por trabajos como (Tu, et al., 2007) y son indicativos del incremento en flujo de electrones de la cadena respiratoria. Así mismo los precursores de aminoácidos y de nucleótidos respaldan los resultados obtenidos por trabajos como (Kuang, et al., 2014) donde se observó que genes codificantes para aminoácidos y de enzimas de la biosíntesis nucleotídica están altamente expresados en dicha fase. Conjuntamente estos resultados apoyan la idea de que la fase OX es una ventana temporal dedicada a la respiración y producción de energía.
Debido a la dualidad de este *cluster* se encuentran también metabolitos asociados a la fase RB. Entre ellos se observan moléculas relacionadas con el aumento de la glicólisis como el glicerol-3-fosfato. Estas observaciones, junto con lo descrito por Sánchez-Gaya sobre expresión génica, refuerzan el hecho de que durante la fase RB la célula cambia a un metabolismo más glicolítico que respiratorio.

Finalmente, el último pico del *cluster* marca la entrada en la fase RC. Y es que algunos metabolitos como, el Acetil-CoA o NADPH, han sido descritos como predictores de dicha fase (Tu, et al., 2005). por lo que el aumento en dicho pico podría ser consecuencia de la entrada en la misma.

En cuanto al primer *cluster,* en este se encuentran metabolitos relacionados con la oxidación de lípidos (beta oxidación), como la carnitina. Igual que en los casos anteriores, estos datos se correlaciona a la perfección con los datos de expresión génica, donde se ha descrito un aumento en la expresión de genes relacionados con la oxidación de lípidos y con la rotura de carbohidratos de almacenaje (Muller, et al., 2003).

3.2 PaintOmics

3.2.1 Nuevo Enfoque de Enriquecimiento Funcional

Como se describió en la sección (2.5) se implementó una nueva funcionalidad a PaintOmics, la posibilidad de realizar un enriquecimiento funcional a partir de asociaciones (regulaciones). Aunque el enriquecimiento por gen no es el más adecuado para el análisis de entidades reguladoras, se decidió hacer uso de los tres tipos de métodos sobre los TFs y modificaciones de histonas. De esa forma, se pretendía analizar las diferencias y testar el funcionamiento de la

nueva opción.

Los resultados generales no se vieron afectados ya que de las 116 rutas evaluadas, 17 resultaron significativas para el p-valor combinado de las ómicas. Como es lógico, los análisis individuales para cada una de las ómicas reguladoras por separado sí que mostraron diferencias relevantes.

Concretamente el enriquecimiento por gen no devolvió ninguna ruta en la cual los TFs se vieran significativamente representados. Por otro lado, esta significatividad sí que se pudo observar tanto en el enriquecimiento por *feature* como por asociación, en 20 y 5 casos respectivamente. La razón de estas diferencias se debe a las variaciones en el conteo de las tablas de contingencia, como se muestra en el ejemplo para la ruta *Ribosomas* (**Tabla 10**).

Como se puede apreciar en la **Tabla 10**, el enriquecimiento por gen toma como total 2035. Este número hace referencia al total de genes únicos que son regulados por al menos un TF y que han sido encontrados en alguna de las rutas KEGG por PaintOmics. En el caso de *feature* se puede ver como este número se ha reducido notablemente a 289. El motivo de ello es que esta vez el programa cuenta en número de TFs únicos encontrados en alguno de las rutas. Finalmente, la nueva aplicación lo que toma como unidad es la combinación de los genes y TFs encontrados. En este último caso 70197.

De esta manera, este nuevo enfoque permite el análisis de las interacciones regulatorias siendo mucho más adecuado en los casos que la relación regulador/gen no es 1/1 y cambiando drásticamente los resultados obtenidos. Sin embargo, en los casos en que la relación sea la anteriormente citada, el enriquecimiento por *feature* o por asociación dará los mismos resultados, como ocurre con los datos de variación de histonas.

Finalmente, es importante destacar que aunque no se vean variaciones en el p-valor conjunto de las ómicas, esto es únicamente debido a que los cambios en el enriquecimiento sólo afectan a los TFs por lo al final no tiene un impacto extremo en este valor. Sin embargo, viendo la gran diferencia en el caso de los TFs, la utilización de este nuevo enriquecimiento en más casos derivaría en resultados diferentes y más realistas.

Tabla 10. Tabla de contingencia resumen de PaintOmics para los tres tipos de enriquecimiento disponibles: por gen, *feature* y asociación. Como relevantes se toman aquellos TFs seleccionados como significativos.

Enriquecimiento (TFs)		Relevantes	No Relevantes	Total
Por gen	Encontradas en la ruta estudiada	167	0	**2035**
	No encontradas	1854	14	
Por *feature*	Encontradas en la ruta estudiada	95	154	**289**
	No encontradas	6	34	
Por Asociación	Encontradas en la ruta estudiada	3304	5189	**70197**
	No encontradas	23228	38476	

3.2.2 Resultados del Enriquecimiento Funcional

Los resultados obtenidos tras la aplicación de PaintOmics mostraron un total de 17 rutas KEGG significativas para el p-valor combinado (**Tabla 11**). Lo más interesante de dichas rutas es que si se analiza la naturaleza de las mismas se encuentran relaciones directas con los diferentes procesos que ocurren a lo largo del YMC descritos en la literatura.

Tabla 11. Rutas significativas obtenidas tras la aplicación del análisis de enriquecimiento en PaintOmics junto con el p-valor combinado. En este caso, dichas rutas están ordenadas ascendentemente en función de dicho p-valor.

Pathway name	pValueCom
Ribosome	0,00
Ribosome biogenesis in eukaryotes	0,00
Aminoacyl-tRNA biosynthesis	0,00
RNA polymerase	0,00
Pyrimidine metabolism	0,00
Purine metabolism	0,00
Sulfur metabolism	0,00
Selenocompound metabolism	0,00
Biosynthesis of secondary metabolites	0,00
Biosynthesis of amino acids	0,00
Peroxisome	0,00
Cysteine and methionine metabolism	0,00
Valine, leucine and isoleucine biosynthesis	0,00
beta-Alanine metabolism	0,02
Pyruvate metabolism	0,02
Biosynthesis of unsaturated fatty acids	0,03
2-Oxocarboxylic acid metabolism	0,03

Un ejemplo es el caso de las rutas de biogénesis de ribosomas, transporte de RNA o producción de aminoácidos. Todas ellas son procesos relacionados con la producción de proteínas, una actividad que se sabe ocurre durante la fase OX del YMC (Kuang, et al., 2014) y que explicaría el aumento de ciertos metabolitos como la homocisteína o el uracilo visto en el apartado (3.1). Por otra parte rutas como peroxisoma podrían estar relacionadas al proceso de división en la levadura durante la fase RB. La activación de los peroxisomas podría estar relacionada con un intento de evitar el daño en el DNA durante dicha división (Chen, et al., 2007). O bien de captar los radicales libres producidos por el aumento en la cadena respiratoria durante la fase OX (Kuang, et al., 2014). Finalmente, se pueden observar rutas

relacionadas con el metabolismo de ácidos grasos, tales como la degradación de ácidos grasos o la biosíntesis de ácidos grasos insaturados. Estas últimas se relacionarían con la fase RC donde la célula presenta una mayor oxidación de ácidos grasos y rotura de carbohidratos de almacenaje (Tu, et al., 2005).

Por otra parte, si se ordena la tabla de resultados en función del p-valor correspondiente a la metabolómica se pueden observar diferentes aspectos. En primer lugar no se observó ninguna ruta en la cual dicha ómica estuviese enriquecida significativamente. Esto es lógico pues el número de metabolitos estudiados es muy pequeño (alrededor de 100), siendo poco recomendable el desarrollo de estudios de enriquecimiento con coberturas tan pobres. Pese a ello, las rutas más significativas siguen manteniendo el sentido biológico del YMC. Y es que, como se puede ver en la **Tabla 12**, también se encuentran algunas relacionadas con la ya descrita producción de aminoácidos, como la biosíntesis de lisina o arginina.

Resumiendo, los resultados presentados por PaintOmics tienen una alta correlación con los procesos principales desarrollados en el YMC. Además, incluso con la adición de la información metabolómica, estos mismos apoyan las conclusiones obtenidas por Sánchez-Gaya en su trabajo de integración.

Tabla 12. Rutas más significativas obtenidas tras la aplicación del análisis de enriquecimiento en PaintOmics junto con el p-valor combinado y el p-valor del enriquecimiento para metabolómica. En este caso, dichas rutas están ordenadas ascendentemente en función del p-valor del enriquecimiento por metabolito.

Pathway name	Metabol	pValueCom
Lysine biosynthesis	0,04	0,24
2-Oxocarboxylic acid metabolism	0,10	0,03
Lysine degradation	0,12	0,59
Arginine biosynthesis	0,12	0,68
Biosynthesis of amino acids	0,13	0,00
Cyanoamino acid metabolism	0,20	0,70

3.3 Regulación del metabolismo: MORE

Se utilizó, en primer lugar, el método MORE para generar un modelo de regresión para cada metabolito en que los predictores del modelo eran los genes. De los 54 metabolitos estudiados, para 47 se obtuvo un modelo significativo con un R^2 superior a 0.7 (**Figura 17**). Por otra parte, del total de genes incluidos en los datos de RNA-Seq (2552), únicamente 59 resultaron ser reguladores significativos en alguno de estos modelos. Finalmente, de acuerdo con estos resultados, la mayoría de metabolitos tienen 1 gen regulador (12 casos), o 2 (28 casos) genes reguladores como variables explicativas significativas (**Figura 17**). Esto se explica por el hecho de que el mínimo número de grados de libertad residual seleccionado fue de 10 y dado que tenemos 14 grados de libertad totales, se permite tener como mucho 4 reguladores significativos.

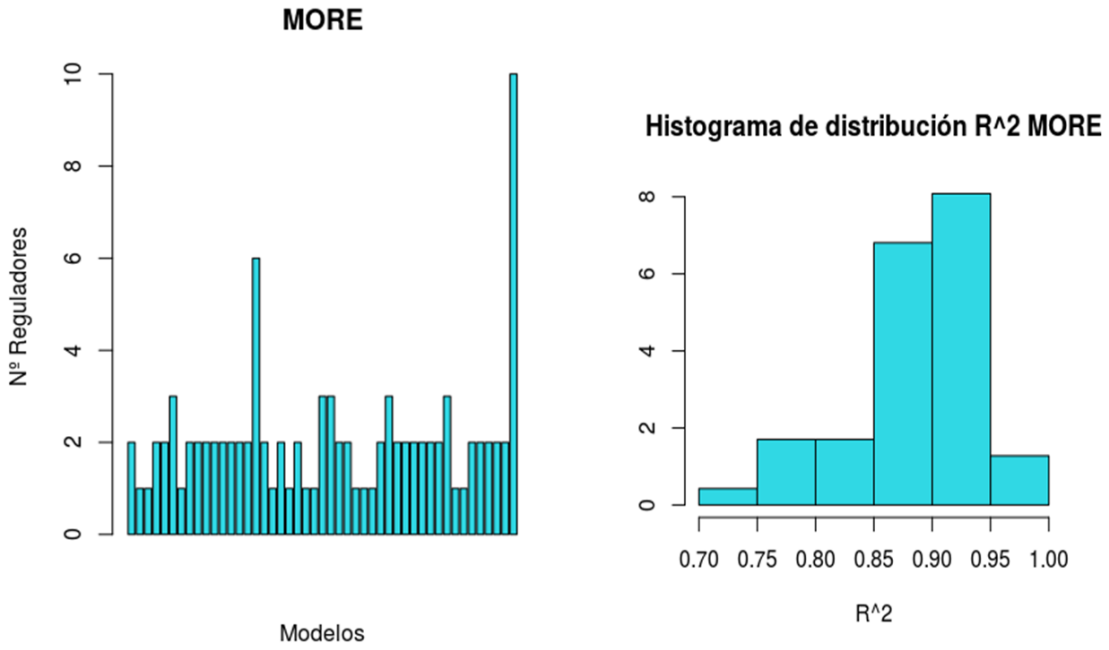


Figura 17. A la izquierda diagrama de barras con el número óptimo de componentes por modelo válido (47). A la derecha, histograma con la distribución del R^2 estos modelos. Estos datos corresponden a los modelos de MORE.

Para acabar, cabe destacar que de las 99 regulaciones significativas, 52 mostraban un valor positivo para su coeficiente de regresión (β) indicando una regulación positiva sobre el metabolito. Por otra parte, 47 eran negativos, sugiriendo una regulación negativa sobre el mismo.

La interpretación biológica de los resultados revela información muy interesante. Por un lado se pueden encontrar algunos metabolitos precursores de nucleótidos, como uracilo y IMP. Estos se ven correlacionados positivamente con RAX2 (**Figura 18**). Dado que la proteína RAX2 participa en la división bipolar de la levadura (Chen, et al., 2000) es comprensible que la producción de nucleótidos se vea aumentada durante dicho proceso. Del mismo modo, como se comentó en el apartado (3.1), estos dos metabolitos se encontraron formando parte del *cluster* 2 por lo que, en base a la bibliografía, se asociaron a la fase OX (Tu, et al., 2005). Sin embargo, dado que la división celular se produce entrada la fase RB podría ser que el aumento de estos metabolitos, así como el inicio en la expresión de RAX2, se produzcan a lo largo de la OX e inicio de RB para así preparar a la levadura para la futura división celular.

Por otra parte también se encuentran metabolitos asociados a la fase OX, como el precursor de aminoácidos homoserina y el intermediario del ciclo del ácido carboxílico el isocitrato. En los modelos de dichos metabolitos aparece una correlación negativa con el factor de transcripción BDF1, en la **Figura 19** se puede ver caso concreto del isocitrato. Según se describe en la bibliografía, esta proteína está asociada a procesos como la reparación del DNA, transcripción, el remodelado de la cromatina y es indispensable para la división celular (Chua, et al., 1995). Conociendo estos datos, es lógico pensar que dicha correlación negativa podría deberse a un paso entre la fase OX a la RB. Dado que durante este cambio, la célula disminuye la producción de aminoácidos aumentando el flujo glicolítico y iniciando la división celular (Tu, et al., 2007). Una prueba más de esto viene de la mano del segundo regulador encontrado en el modelo del isocitrato, concretamente IQG1 (**Figura 19**). Y es que esta proteína participa en la formación del anillo de actina durante la división (Epp y Chant, 1997) por lo que de nuevo la

anterior hipótesis podría ser válida.

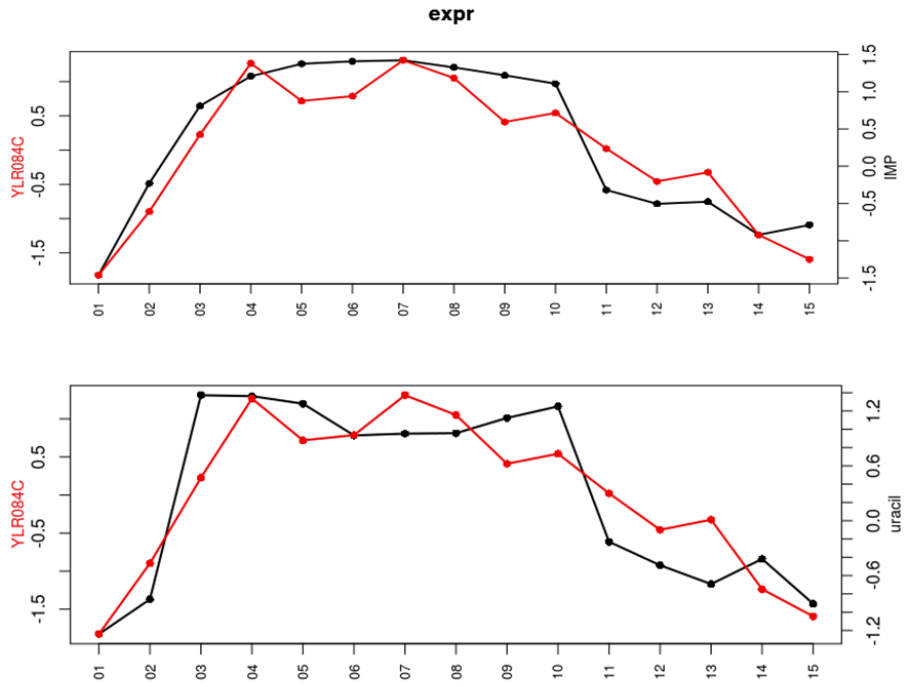


Figura 18. Patrón de Expresión de *RAX2* (YLR084C) junto con el perfil del metabolito IMP (arriba) o uracilp (bajo) a lo largo del tiempo.

Finalmente un ejemplo de un metabolito relacionado con la fase RC es el Acetil-CoA. El modelo de MORE muestra dos genes como variables explicativas, *EPL1* y *NNT1*. En el primero de ellos se observa una relación negativa con el metabolito. Esta podría ser producto de la función de dicho gen. Y es que *EPL1* es un componente indispensable del complejo H4/H2A acetil-transferasa (Stankunas, et al., 1998). Como describió Sánchez-Gaya en su trabajo, la acetilación H4K5ac en combinación con MIG2 parecen estar implicados en el metabolismo de carbohidratos. Sin embargo, este proceso se ve disminuído a lo largo de la fase RC momento en el cual predomina el metabolismo lipídico y que se caracteriza por un aumento en Acetil-CoA. Este paso a RC podría ser el motivo de dicha relación negativa. Por otro lado, la relación positiva con *NNT1* no se explica tan fácilmente. Pese a ello, algunos trabajos han relacionado esta proteína con un aumento en la viabilidad de la células en periodos de restricción calórica (Anderson, et al., 2003) dado que en esta última fase, RC, la restricción calórica de la levadura es mayor por lo que podría ser la razón de esta relación positiva entre la proteína y el Acetil-CoA.

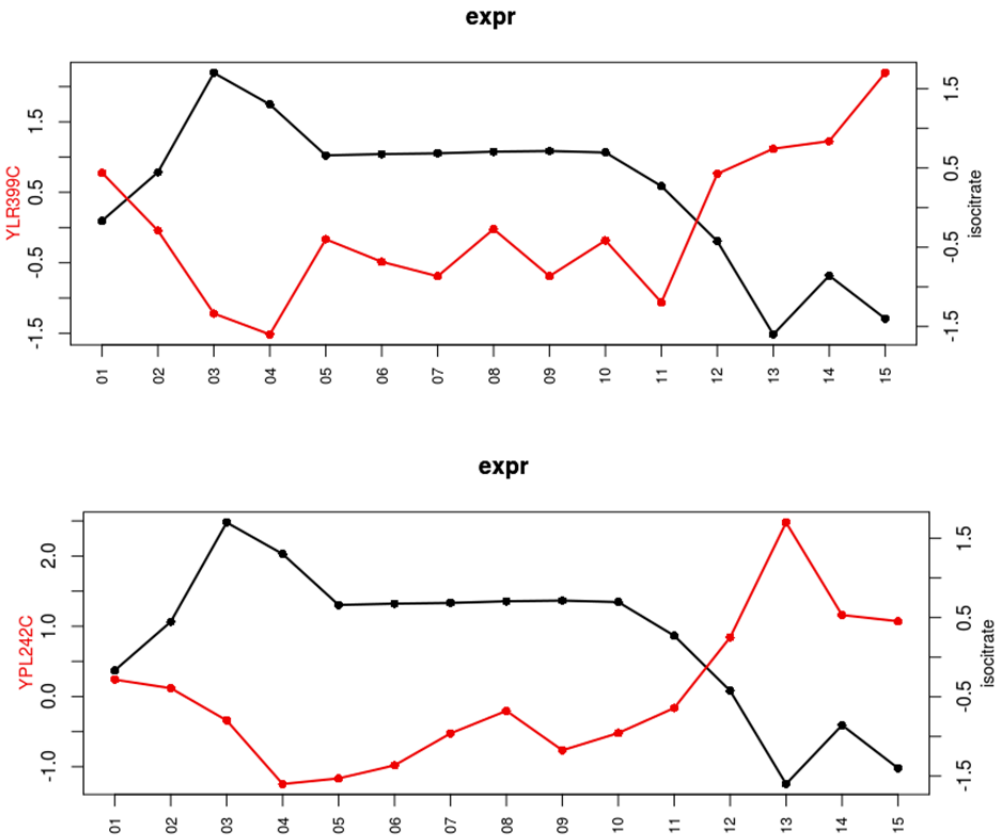


Figura 19. Patrón de Expresión de *BDF1* (YLR399C) junto con el perfil del Isocitrato. Abajo se muestra el patrón de expresión de *IQGL1* junto al mismo metabolito anterior.

3.4 Regulación del metabolismo: PLS

Como alternativa al método MORE, se aplicó la metodología PLS para estudiar la regulación del metabolismo. Se generó un modelo PLS para cada metabolito siendo este la variable respuesta y los 2552 genes los predictores.

El primer paso del análisis PLS consistió en determinar el número óptimo de componentes para cada uno de los 54 modelos. Dado que el objetivo de estos modelos era más tratar de explicar la relación entre genes y metabolitos que predecir nuevas observaciones, se optó por elegir el número de componentes que optimizaran el valor de R^2 del modelo. Para ello, tal y como se describió en el apartado 2.6.2, se utilizó el criterio SIMCA-P sobre este estadístico.

A partir de este criterio se encontró que el número de componentes óptimas para cada modelo estaba dentro del rango de entre una a cuatro componentes (**Figura 20**). Si bien se observó que en la mayoría de los casos, 25 de 54, únicamente con 2 componentes se alcanzaba el nivel óptimo de R^2 (**Figura 20**).

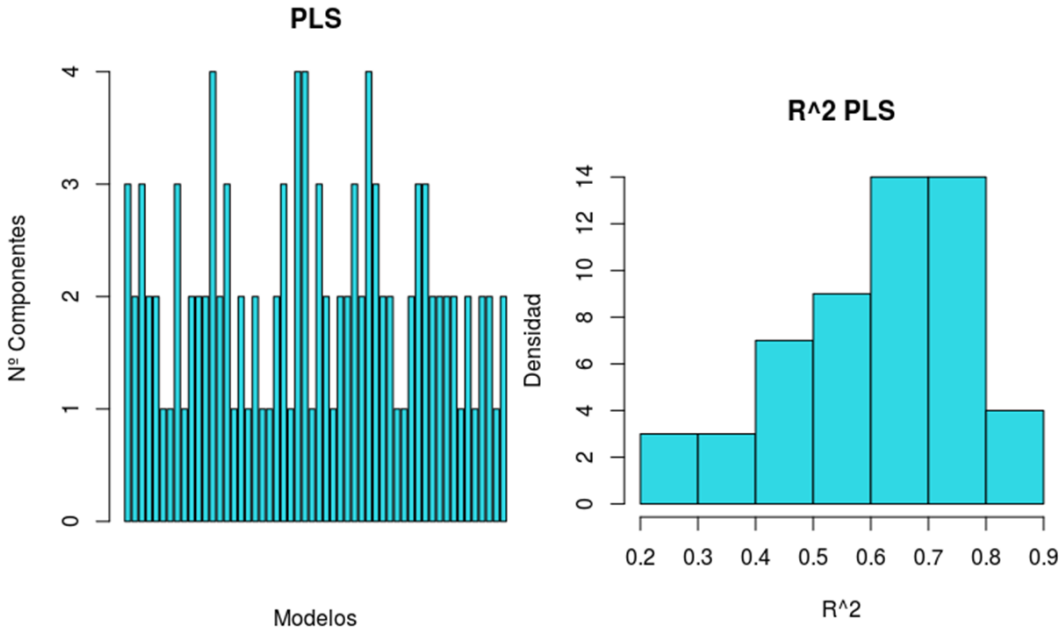


Figura 20. A la izquierda diagrama de barras con el número óptimo de componentes por modelo. A la derecha, histograma con la distribución del R^2 de los modelos. Estos datos corresponden a los modelos antes de la selección de variables.

Tras ello, se determinó el número óptimo de variables (genes) a retener en cada modelo mediante el método de selección de variables del *sparse*-PLS seguido del criterio SIMCA-P sobre el R^2 (**Figura 21**). Se observó que el número medio adecuado de variables por componente era de entre diez y veinte en la mayor parte de los casos (30 de 54 modelos, **Figura 22**).

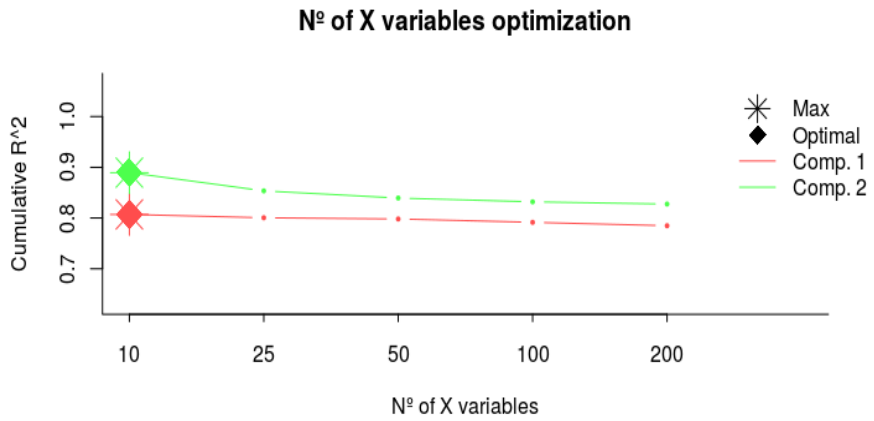


Figura 21. R^2 acumulado en función del número de variables por cada componente (Comp.1 y Comp.2) para el metabolito Acetil-CoA. Cada recta queda marcada por el número de componentes que genera el mayor valor de R^2 (estrella) y el número óptimo de componentes siguiendo el criterio SIMCA-P (rombo).

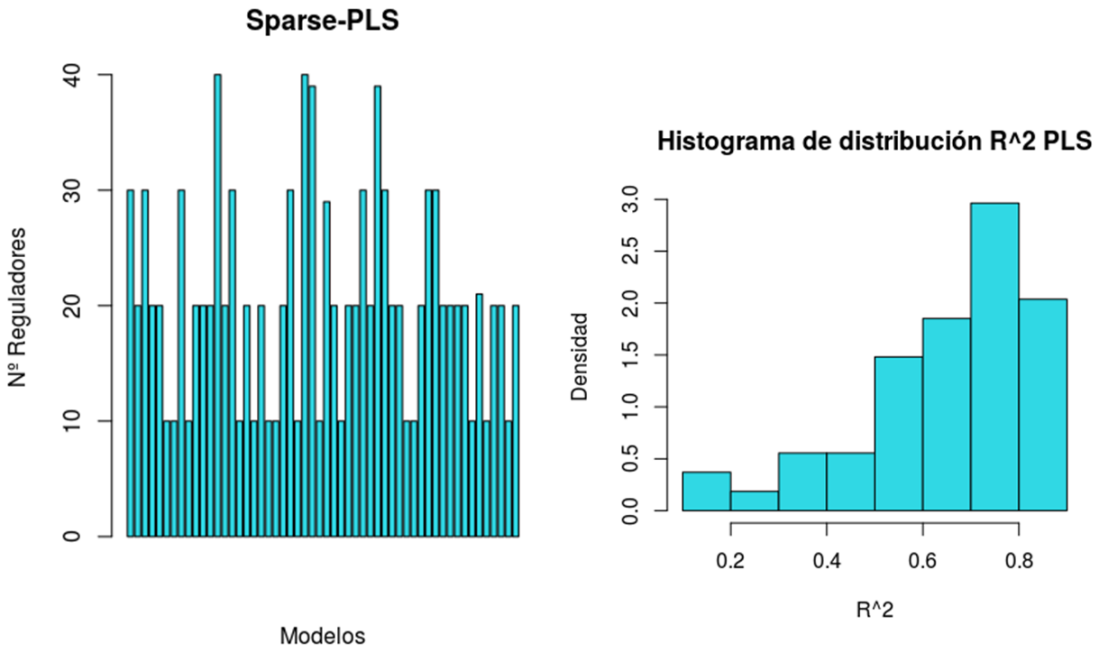


Figura 22. A la izquierda diagrama de barras con el número óptimo de reguladores por modelo. A la derecha, histograma con la distribución del R^2 de los modelos. Estos datos corresponden a los modelos PLS-*sparse*.

Con el objetivo de analizar la mejora, o empeoramiento, de la reducción de variables por componente, se compararon los valores de los estadísticos: R^2, Q^2 y MSEP (Mean Squared Prediction Error, de sus siglas en inglés) (**Figura 23**).

Lo que se observó es que en el caso de R^2 la reducción de variables mejoraba este estadístico en la mayor parte de los casos mientras que en el resto o bien se mantenía casi equivalente o empeoraba muy poco. En cuanto al Q^2 si que pudo apreciar un empeoramiento del mismo por la reducción de variables. Sin embargo, dado que el interés principal del modelo no era la predicción de nuevas observaciones si no la interpretación de las relaciones metabolito-gen, no se tuvo en cuenta este empeoramiento. Finalmente para el caso de MSEP se vio, en partes iguales, una ligera mejora en algunos modelos y en otros un pequeño empeoramiento. Sea como sea, dado el objetivo del estudio, se utilizó el R^2 como criterio principal para la comparación entre modelos.

Concretamente se pudo ver que en 27 de 54 modelos la reducción del número de variables afectó positivamente al este estadístico. Un ejemplo de esta mejora se pudo ver en el metabolito SAH donde se incrementó de un valor de 0.68 a 0.82. Por otro lado, como se comentaba en el párrafo anterior, en los modelos donde se observó un empeoramiento del R^2, este era muy pequeño. Un ejemplo se puede ver con el metabolito tiamina donde la reducción de variables produjo un decrecimiento del R^2 del 0.59 al 0.55.

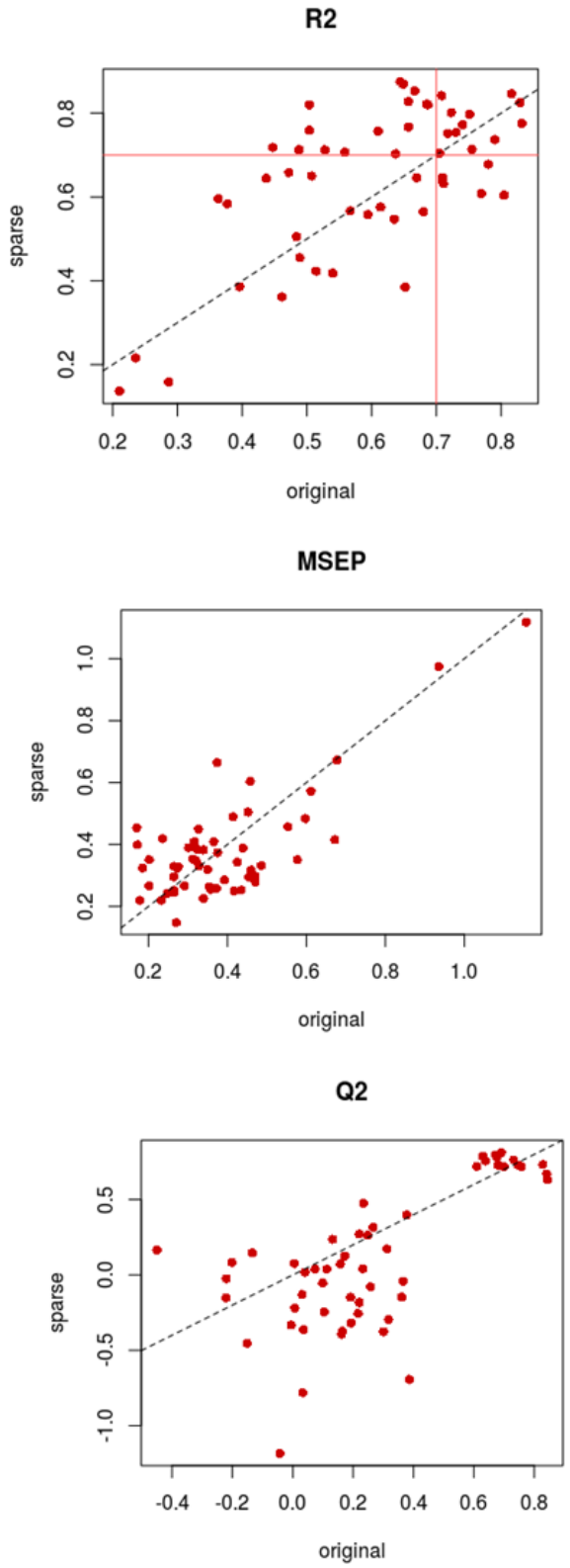


Figura 23. Se comparan los estadísticos R^2, MSEP y Q^2 entre los modelos PLS originales y su variante *sparse*. La línea discontinua delimita la mejora/empeoramiento de dicho estadístico cuando se produce la selección de variables. En el caso concreto de R^2, las lineas rojas señalan el umbral de 0.7 seleccionado para marcar la "validez" del modelo.

Posteriormente, para la selección de los modelos "válidos" se utilizó el mismo criterio que en los modelos MORE, un R^2 superior a 0.70 (**Figura 23**). El filtrado mediante este criterio estableció que de los 54 modelos únicamente 27 fueran considerados como válidos. Esto resultó en un número muy pequeño de modelos por lo que se decidió realizar un estudio de correlaciones entre cada metabolito y los genes seleccionados en su modelo *sparse* correspondiente.

Este estudio reveló que en algunos casos los genes incluidos en el modelo no presentaban un elevado valor de correlación con el metabolito (**Figura 24**).Por ello, se decidió seleccionar aquellos genes cuya correlación superaba el 0.8 en valor absoluto. Esto resultó en la eliminación de uno de los modelos al completo quedando un total de 53 metabolitos con genes reguladores. Por otro lado, se produjo un decrecimiento en el número de genes únicos incluidos en alguno de los modelos pasando de 601 a 305.

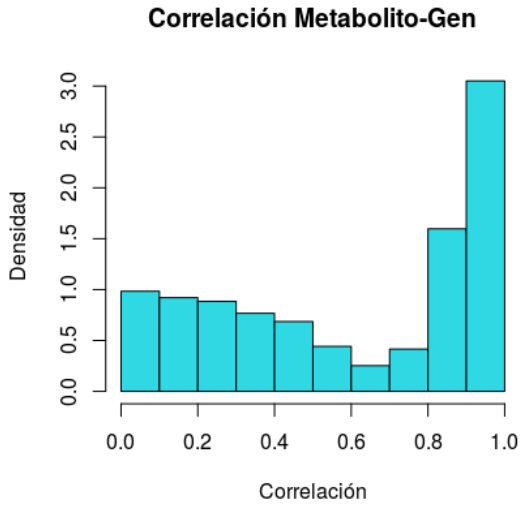


Figura 24. Histograma de la distribución de la correlación entre metabolito y gen. Los valores de la correlación están en valor absoluto. Estos datos corresponden al modelo *Sparse*-PLS.

Sobre estos 305 genes se realizó la interpretación biológica de algunos metabolitos. Y es que de nuevo, al igual que ocurrió con el MORE, se pudieron ver resultados muy interesantes.

Un ejemplo es el caso del metabolito IMP, este metabolito, tal y como se describió en el apartado 3.3, es un precursor de nucleótidos que se asocia a la fase OX del YMC. En el modelo de MORE se encontró que RAX2 correlacionaba positivamente con dicho metabolito y se trató de argumentar esta relación. Y es que, RAX2 participa en la división bipolar de la levadura (Chen, et al., 2000), un proceso que está ligado a la fase RB por encima de la fase OX. Es por ello que resultaba algo extraña dicha correlación positiva. Sin embargo, en el modelo de PLS no encontramos a RAX2 y si que se encuentran otras proteínas que encajan mucho mejor con los actuales conocimientos del YMC (**Figura 25**). Por ejemplo, una de ellas es PAU3 (YCR104W), esta proteína regula la fermentación alcohólica en levadura y que se conoce es fuertemente inhibida por la presencia de oxígeno en el ambiente (Rachidi, et al., 2000). Dado que IMP aparece durante la fase OX donde el metabolismo aeróbico es predominante es lógico encontrar una elevada correlación negativa (en torno al -0.94) con PAU3. Del mismo modo, se encuentran otras proteínas con una correlación muy positiva, es el caso de MRS3 (YJL133W). Esta proteína se relaciona con el transporte del ión Fe2+ en la mitocondria en situaciones donde existe una limitación de dicho elemento (Mühlenhoff, et al., 2003). Debido a la situación

oxidativa de la célula en la fase OX, la cadena de transporte mitocondrial debería de estar muy activa siendo necesaria la adquisición de hierro para la formación de las proteínas que forman parte de la misma. Adicionalmente, el hecho de que el YMC aparezca bajo condiciones de limitación de nutrientes apoya el hecho de que MRS3 aumente en la fase OX junto a metabolitos como IMP.

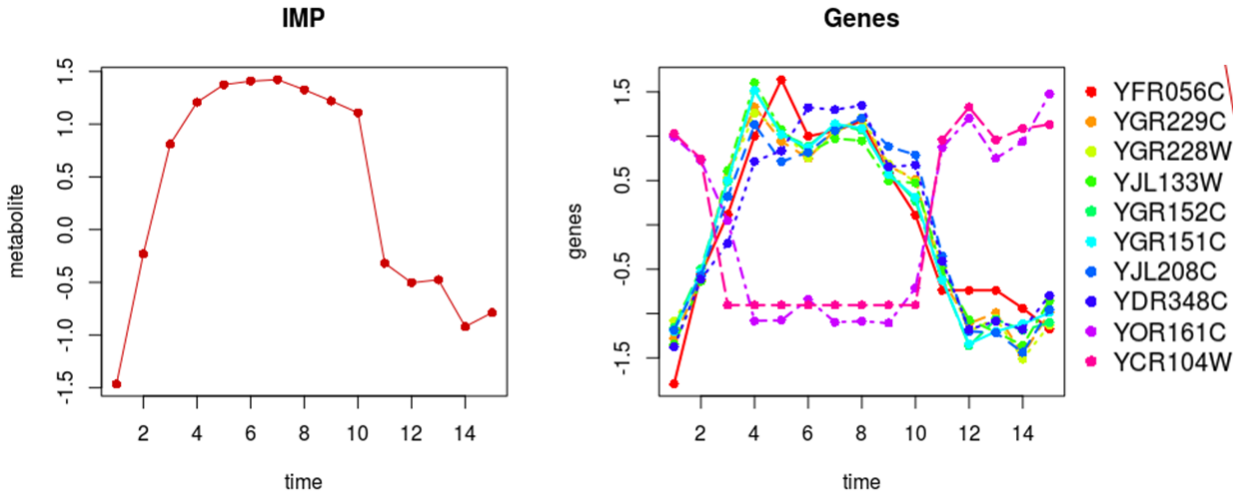


Figura 25. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito IMP. A la derecha el perfil de los genes asociados al mismo por el modelo *Sparse*-PLS tras el filtro de correlación.

Otro ejemplo de metabolito es el isocitrato, en este caso sí que incluye todos los genes encontrados en el modelo MORE, concretamente *BDF1* (YLR399C) y *IQG1* (YPL242C) (**Figura 26**). Del mismo modo que ocurría con el primer modelo, se encuentra una correlación negativa con dichas proteínas. Sin embargo, lo más interesante aparece por la inclusión de nueva información por parte del modelo PLS. Y es que en este caso aparecen nuevas proteínas como LYS9 que forma parte de la ruta de biosíntesis de la lisina (Borell, et al., 1984). Esta proteína presenta una clara correlación positiva con el isocitrato y que se debe a que tanto la producción de aminoácidos como el ciclo carboxílico se incrementan a lo largo de la fase OX del YMC (Tu, et al., 2007).

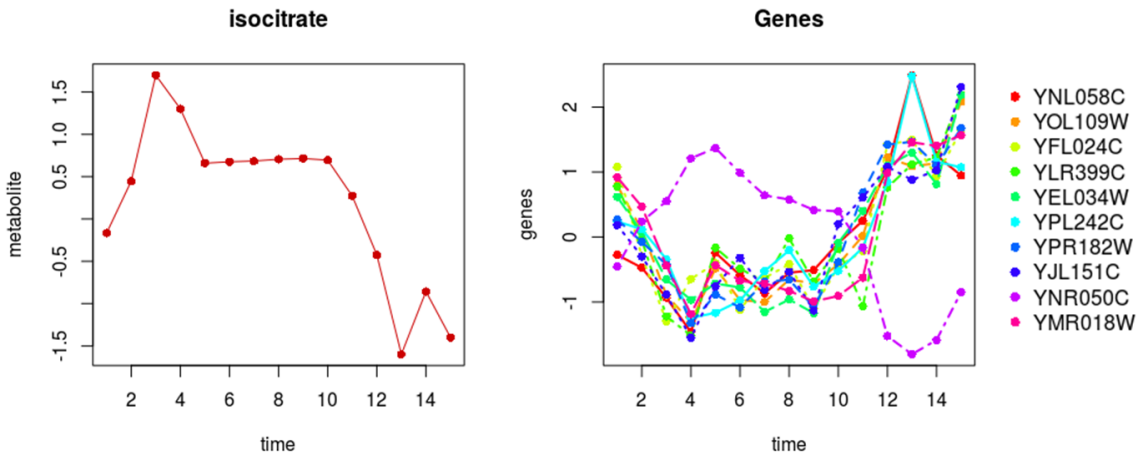


Figura 26. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito isocitrato. A la derecha el perfil de los genes asociados al mismo por el modelo *Sparse*-PLS tras el filtro de correlación.

Finalmente un último ejemplo de modelo relacionado con YMC es la carnitina. Esta se ve incrementada a lo largo de la fase RC (Tu, et al., 2007). Una de las relaciones más destacables encontradas en este modelo es la relación con la proteína PAU3 (YCR104W) (**Figura 27**). Y es que, a diferencia de lo que se explicaba con el metabolito IMP, la correlación de la carnitina con PAU3 es totalmente positiva. Esto se explica ya que la carnitina es un metabolito que aparece incrementado a lo largo de la fase RC, una fase en la que la célula presenta un metabolismo reductivo y no oxidativo. Consecuentemente, la ausencia de oxígeno característica de la RC hace que PAU3 no esté inhibida, al contrario que lo que ocurría con IMP que se encuentra en una fase oxidativa como es la OX (**Figura 25**).

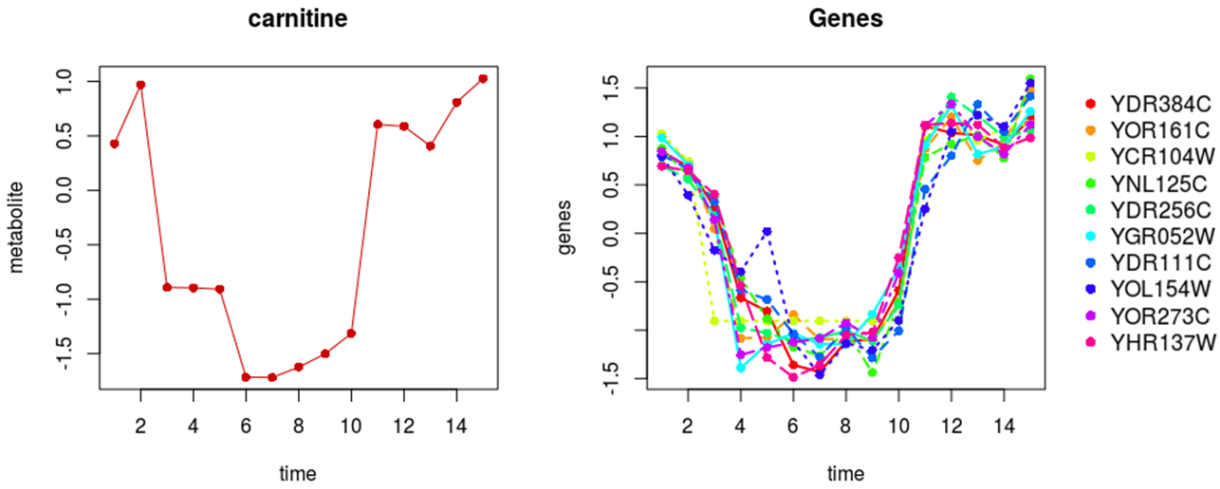


Figura 27. Perfiles metabolito/genes a lo largo de los 15 puntos temporales. A la izquierda se muestra el perfil del metabolito carnitina. A la derecha el perfil de los genes asociados al mismo por el modelo *Sparse*-PLS tras el filtro de correlación.

3.5 Comparación modelos MORE y PLS

Como se ha podido observar a lo largo de los apartados 3.2 y 3.3 los modelos obtenidos por la metodología MORE y PLS presentan algunas diferencias. La primera de ellas guarda relación con el número de variables explicativas que forman parte de cada uno de los modelos. En el caso de MORE, dado que está basado en un modelo de regresión lineal tradicional el número de variables explicativas es bastante limitado y depende directamente del número de observaciones. Esto es un problema sobretodo en estudios ómicos donde el número de observaciones rara vez es elevado, en este caso 15, por lo que el número de grados de libertad totales restringen significativamente el número de variables explicativas por modelo. Esto no ocurre con el PLS ya que al ser un método multivariante está preparado para ser aplicado en estos casos. Como consecuencia los modelos obtenidos por MORE presentaban una media de dos variables explicativas mientras que en el PLS, tras la selección de variables, se observaron 10 en promedio. No solo eso sino que el número de genes distintos incluidos en los 47modelos MORE fueron 59 frente a los 305 delos53 modelos PLS.

En cuanto a la comparación entre modelos para el mismo metabolito, para un 60% de los 47 modelos MORE los genes significativos no coinciden con los seleccionados por el correspondiente modelo PLS, mientras que para un 28% el PLS selecciona los mismos genes que MORE y alguno más adicional. Para el resto de metabolitos, los genes seleccionados por

MORE y PLS coinciden parcialmente.

Se estudió también la correlación entre los genes seleccionados y el metabolito para cada uno de los modelos (en el caso del *sparse* PLS antes del filtro por correlación). Se puede observar en la gráfica que los genes seleccionados por MORE tienen una correlación mucho mayor con su metabolito que los derivados del PLS, tanto en mínimo, como en máximo, mediana y media (**Figura 28**).

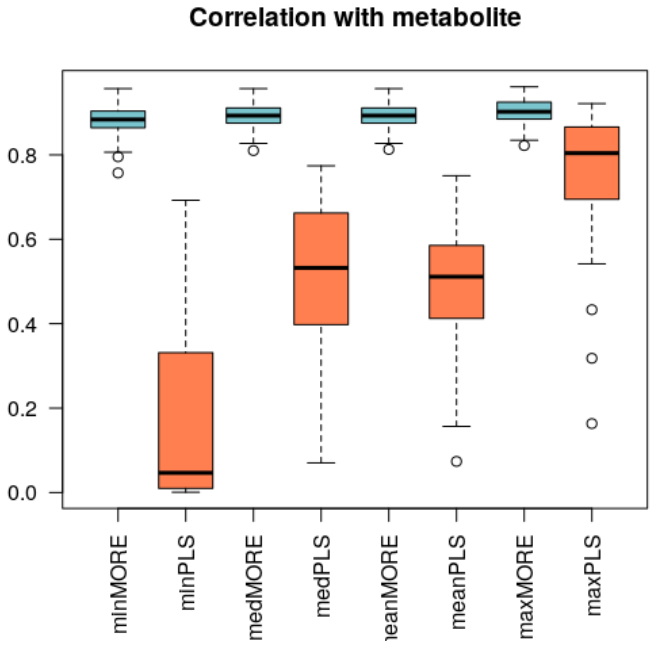


Figura 28. Diagrama de cajas de la correlación entre los modelos y sus metabolitos. Los modelos comparados son: MORE (azul) y PLS-*sparse* (antes del filtro por correlación, rojo). Se compara la correlación en mínimo (min), máximo (max), media (mean) y mediana (med).

A la vista de estos resultados, cabe preguntarse: ¿Por qué PLS y MORE no comparten un mayor número de genes? ¿Por qué tanto MORE como PLS no detectan algunos genes altamente correlacionados con el metabolito? Para dar respuesta a estas cuestiones, se estudiaron en detalle algunos de los modelos obtenidos.

En cuanto a MORE, conviene recordar que era la primera vez que se probaba utilizando como reguladores potenciales todos los reguladores disponibles, en lugar de solo una parte de ellos definida en bases de datos, literatura, experimentalmente, etc. Al analizar, pues, algunos modelos MORE se detectó un problema que, si bien podría darse con un menor número de reguladores potenciales, se incrementa enormemente cuando el número de dichos reguladores es tan elevado. Este problema está relacionado con el llamado filtro de multicolinealidad que aplica MORE. Este filtro consiste en determinar qué reguladores están altamente correlacionados entre sí (en nuestro caso con correlación en valor absoluto mayor a 0.95) y elegir entre ellos (de forma aleatoria) un representante, que será el único que se incluya en la ecuación inicial del modelo de regresión. En este estudio, por tanto, se analiza la correlación entre los más de 2000 genes que se analizan. Es posible que el gen G1 esté muy correlacionado con el gen G2 y este a su vez con el gen G3 y así sucesivamente hasta llegar al gen Gn. Todos ellos se considerarían un grupo de genes altamente correlacionado pero es posible que el gen G1 y el Gn tengan una correlación muy baja y que por tanto no tenga sentido que sean representados por el mismo gen representante. Como se comentaba anteriormente, esta situación se dará con mayor probabilidad cuanto mayor sea n y esto a su vez cuanto mayor sea el número de genes considerados como reguladores potenciales del

metabolito. Así pues, este trabajo ha servido, entre otras cosas, para poner de manifiesto este problema que será corregido en la futura versión del método MORE por los desarrolladores del paquete. Además, por consiguiente, esto explica por qué MORE no detecta algunos genes como significativos que sí detecta el PLS. En todos los casos analizados, estos genes formaban parte de un grupo de genes altamente correlacionados pero desafortunadamente se seleccionó un representante con baja correlación con el metabolito y por ello no pasaron el filtro de selección de variables de MORE.

Por otra parte, faltaría entender porque no siempre el *sparse* PLS selecciona como mínimo los mismos genes que MORE que, como se ha visto (**Figura 28**) presentan una alta correlación con el metabolito. Se podría hipotetizar en este caso que el modelo PLS tiende a seleccionar variables que, no solo correlacionan con la variable respuesta, sino que también correlacionan con otras variables explicativas, y que aquellos genes con un comportamiento más "independiente" no son considerados. Esto cuadraría con el hecho de que esos genes no han formado parte de un grupo de genes altamente correlacionados en el algoritmo MORE. Además, el modelo PLS selecciona genes con baja correlación con la variable respuesta y esto puede ser debido a que están correlacionados con algunos de los genes que sí que coexpresan con el metabolito y por tanto tienen alto *loading* en esa componente.

Así pues, como se ha visto, tanto MORE como PLS tienen sus ventajas e inconvenientes, no se podría decir que un modelo es mejor que otro en todos los casos. En ambos se requiere refinar mejor el proceso de selección de variables y, de hecho, el grupo de Genómica de la Expresión Génica está trabajando ahora mismo en ello.

4. Conclusiones

A lo largo de este trabajo pude extraer diferentes conclusiones. La primera de todas es que existen una gran variedad de metodologías estadísticas para la integración de datos ómicos. Siendo el problema no tanto la disponibilidad si no la adecuación del modelo a los datos y objetivos del investigador. Así mismo, también es importante la comparación de estos modelos pues el área de integración de datos es una especialización relativamente nueva y aún hace falta mucho esfuerzo para llegar a entender cuando se debe aplicar un tipo de modelo u otro. De hecho, como se puede haber visto a lo largo de este trabajo, se encontraron diferencias inesperadas entre los modelos MORE y PLS que pueden afectar directamente a la interpretación biológica de los resultados. Es por ello que, el primer punto a tratar en un futuro debería ser el análisis comparativo de diferentes estrategias estadísticas.

Por otra parte, también pude extraer conclusiones biológicas acerca del YMC. Viendo algunas relaciones que pueden ser explicadas de forma "sencilla" y otras que nos eran desconocidas o no habían sido descritas. Esto es una de las virtudes, y en ocasiones inconveniente, de la biología de sistemas que permite encontrar interacciones no conocidas. El análisis más profundo de estas interacciones podría ser un camino a tomar en el futuro

Finalmente, aunque no incluidas dentro del trabajo, también aprendí a que no siempre la idea inicial de un proyecto permanece inalterada y que la capacidad de adaptación es muy importante. Esto me ocurrió cuando intentamos aplicar modelos como, redes Bayesianas o el PLS-PM (PLS-Path Modeling, de sus siglas en inglés). Y es que la dificultad que encontramos en la aplicación de dichos modelos hizo que tuvieran que ser desechados y, consecuentemente, no incluidos en el trabajo.

Para acabar, algo que me gustaría realizar en un futuro y que completaría la presente tesis de máster sería el análisis de la relación entre las modificaciones de histonas y el metaboloma. Para ello uno de los métodos estadísticos posibles sería el O2-PLS.

5. Glosario

EM	Expectation–Maximization algorithm
GCxGC-TOFMS	2D Gas Chromatography - Time-of-Flight Mass Spectrometer
GEO	Gene Expression Omnibus
GLM	Generalized Linear Model
Gtf	Gene Transfer Format
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS/MS	Liquid Chromatography–Mass/Mass spectrometry
LASSO	Least Absolute Shrinkage and Selection Operator
MICE	Multivariate Imputation by chained Equations
MORE	Multi-Omics Regulation R Methodology
MSEP	Mean Squared Prediction Error
OX	Oxidative
Pb	Base Pairs
PCA	Principal Component Analysis
PEC	Pruebas de Evaluación Continua
PLS	Partial Least Squares regression
O2-PLS	Two-Way Orthogonal Partial Least Squares
PLS-PM	Partial Least Squares Path Modeling
RB	Reductive Building
RC	Reductive Charging
TF	Transcription Factor
TSS	Transcription Start Sites
YMC	Yeast Metabolic Cycle

6. Bibliografía

Anderson, R. M., Bitterman, K. J., Wood, J. G., Medvedik, O., & Sinclair, D. A. (2003). Nicotinamide and PNC1 govern lifespan extension by calorie restriction in Saccharomyces cerevisiae. *Nature*, *423*(6936), 181.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Biagioni, D. J., Astling, D. P., Graf, P., & Davis, M. F. (2011). Orthogonal projection to latent structures solution properties for chemometrics and systems biology data. *Journal of Chemometrics*, *25*(9), 514-525.

Borell, C. W., Urrestarazu, L. A., & Bhattacharjee, J. K. (1984). Two unlinked lysine genes (LYS9 and LYS14) are required for the synthesis of saccharopine reductase in Saccharomyces cerevisiae. *Journal of bacteriology*, *159*(1), 429-432.

Bro, R. (1996). Multiway calibration. multilinear pls. *Journal of chemometrics*, *10*(1), 47-61.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

Cavill R, Jennen D, Kleinjans J, Briedé JJ. (2016). Transcriptomic and metabolomic data integration. Briefings in bioinformatics, 17(5): 891 - 901. doi: 10.1093/bib/bbv090.

Chen, T., Hiroko, T., Chaudhuri, A., Inose, F., Lord, M., Tanaka, S., ... & Fujita, A. (2000). Multigenerational cortical inheritance of the Rax2 protein in orienting polarity and division in yeast. *Science*, *290*(5498), 1975-1978.

Chen, Z., Odstrcil, E. A., Tu, B. P., & McKnight, S. L. (2007). Restriction of DNA replication to the reductive phase of the metabolic cycle protects genome integrity. *Science*, *316*(5833), 1916-1919.

Conesa A, Nueda MJ, Ferrer A, Talón M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics 1;22: 1096 - 102.

Chua, P., & Roeder, G. S. (1995). Bdf1, a yeast chromosomal protein required for sporulation. *Molecular and cellular biology*, *15*(7), 3685-3696.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, *4*(9), R60.

Draper, N. R., & Smith, H. (1998). Stepwise regression. *Applied Regression Analysis*, *307*, 312.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis."*Bioinformatics*, 21, 3439–3440.

Durinck S, Spellman P, Birney E, Huber W (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nature Protocols*, 4, 1184–1191.

Epp, J. A., & Chant, J. (1997). An IQGAP-related protein controls actin-ring formation and cytokinesis in yeast. *Current Biology*, *7*(12), 921-929.

Faraway, J. J. (2005). Extending the linear model with R (Texts in Statistical Science).

Fearn, T. (2000). On orthogonal signal correction. *Chemometrics and intelligent laboratory systems*, *50*(1), 47-52.

Furió-Tarí, P., Conesa, A., & Tarazona, S. (2016). RGmatch: matching genomic regions to proximal genes in omics data integration. *BMC bioinformatics*, *17*(15), 427.

Gustin, M. C., Albertyn, J., Alexander, M., & Davenport, K. (1998). MAP Kinase Pathways in the YeastSaccharomyces cerevisiae. *Microbiology and Molecular biology reviews*, *62*(4), 1264-1300.

Hernández-de-Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furio-Tari, 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*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Horgan R and Kenny L. (2011). "Omic" technologies: genomics, transcriptomics, proteomics and metabolomics. Obstetrics Gynecology, 13: 189 - 195.

James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1–47. URL http://www.jstatsoft.org/v45/i07/.

Kuang Z, et al.(2014). High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. Nature Structural and Molecular Biology, 21(10): 854 - 63. doi: 10.1038/nsmb.2881.

Lê Cao, K. A., González, I., & Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, *25*(21), 2855-2856.

M.C. Teixeira, P.T. Monteiro, M. Palma, C. Costa, C.P. Godinho, P. Pais, M. Cavalheiro, M. Antunes, A. Lemos, T. Pedreira, I. Sá-Correia (2018). YEASTRACT, an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucl. Acids Res., 46 (Issue D1): D348-D353.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., & Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, *17*(4), 628-641.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. *The R Journal*, *9*(1), 207-218.

Mühlenhoff, U., Stadler, J. A., Richhardt, N., Seubert, A., Eickhorst, T., Schweyen, R. J., ... & Wiesenberger, G. (2003). A specific role of the yeast mitochondrial carriers MRS3/4p in mitochondrial iron acquisition under iron-limiting conditions. *Journal of Biological Chemistry*.

Müller, D., Exler, S., Aguilera-Vázquez, L., Guerrero-Martín, E., & Reuss, M. (2003). Cyclic AMP mediates the cell cycle dynamics of energy metabolism in Saccharomyces cerevisiae. *Yeast*, *20*(4), 351-367.

Nelder, J. A., & Baker, R. J. (2004). Generalized linear models. *Encyclopedia of statistical sciences*, *4*.

Nielsen J. (2003). It is all about metabolic fluxes. Journal of Bacteriology, 185:
7031 - 7035.

Rachidi, N., Martinez, M. J., Barre, P., & Blondin, B. (2000). Saccharomyces cerevisiae PAU genes are induced by anaerobiosis. *Molecular microbiology*, *35*(6), 1421-1430.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, *26*(3), 303.

Rohart, F., Gautier, B., Singh, A., & Le Cao, K. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, *13*(11), e1005752.

Sánchez-Gaya V, et al.(2018).Elucidating the regulation of the Yeast Metabolic Cycle through the integration of gene expression and chromatin status. Frontiers in Genetics, (under revision).

Sanchez, G. (2013). PLS path modeling with R. *Berkeley: Trowchez Editions*, *383*.

Sanchez, G., Trinchera, L., Sanchez, M. G., & FactoMineR, S. (2013). Package 'plspm'.

Stankunas, K., Berger, J., Ruse, C., Sinclair, D. A., Randazzo, F., & Brock, H. W. (1998). The enhancer of polycomb gene of Drosophila encodes a chromatin protein conserved in yeast and mammals. *Development*, *125*(20), 4055-4066.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., ... & Sa-Correia, I. (2006).

The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic acids research*, *34*(suppl_1), D446-D451.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series*

B (Methodological), 267-288.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics: A Journal of the Chemometrics Society*, *16*(3), 119-128.

Trygg, J. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *16*(6), 283-293.

Tu BP, Kudlicki A, Rowicka M, McKnight SL. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science, 310(5751): 1152 - 8.

Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures in: David, FN (Hrsg.), Festschrift for J. *Neyman: Research Papers in Statistics, London*.

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent laboratory systems*, *44*(1-2), 175-185.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., ... & Gil, L. (2017). Ensembl 2018. *Nucleic acids research*, *46*(D1), D754-D761.

Zhang, J. D., & Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, *25*(11), 1470-1471.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.