INTEGRACIÓN DE DATOS DE PROTEÓMICA Y METABOLÓMICA: APLICACIÓN AL ESTUDIO DE LA ATEROSCLEROSIS SUBCLÍNICA

# Preprocesado de los datos

## PESA

### Metadatos



Número de observaciones en las cohortes de PESA y su distribución entre controles y casos.



Mann-Whitney U test para determinar el grado de asociación entre los metadatos y la condición caso/control. Representación mediante Boxplot de la distribución de cada variable comparando caso y control en este [enlace](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metadata/PESA/WorkingFiles/MetadataStats.html).

### Proteómica

**Generación y preprocesado de los datos**

Se realizó un experimento de Proteómica no dirigida cuyo objetivo era la cuantificación relativa de las proteínas en cada una de las muestras. Para ello, las muestras pertenecientes a cada cohorte se dividieron en lotes (batches) para su procesado y cuantificación empleando marcaje isobárico (TMT). Para poder comparar las muestras de diferentes lotes, uno de los canales de cada TMT se reservó para marcar un pool de muestras, que será usado como referencia común (denominador en el ratio) en el proceso de cuantificación.

Para la cuantificación se utilizó el paquete SanXoT (WSPP), que permite realizar integraciones scan2péptido2proteína2all. En este caso, la tabla primaria de datos es una matriz SxN (número de scanes x número de muestras) con los valores de intensidad detectado por el equipo en los diferentes canales del TMT correspondiente. El proceso de cuantificación comienza con el cálculo de los log2-ratios de los scanes:

El WSPP calculará el log2-ratio de los péptidos como una media ponderada del log2-ratio de sus scanes. Asimismo, el log2-ratio de las proteínas se obtiene como una media ponderada de los valores obtenidos para sus péptidos. Este proceso de integración se realiza para cada muestra/canal y aplicando un algoritmo iterativo que permite estimar la varianza asociada al proceso de integración. Para poder comparar los valores X de diferentes muestras el SanXoT permite calcular el log2-ratio global de cada muestra en la integración proteína2all. Por tanto, para una proteína q en una muestra e, se calculará el denominado LowerNorm:

Esta operación permite corregir las diferencias en las concentraciones de partida de cada una de las muestras, la desviación sistemática en la señal de los canales… El material suplementario de la publicación del WSPP indica parte de lo expuesto anteriormente:

Diagram, schematic

Description automatically generated

Diagram, schematic

Description automatically generated

En su utilización habitual, el SanXoT calcula los valores de Zq mediante una tipificación del valor Xq:

A picture containing text, clock, watch

Description automatically generated

Esta transformación estandariza los valores de las proteínas por cada una de las muestras (media = 0 y std = 1) y además permite incluir el peso de la proteína en el valor final. Esta aproximación difiere de la aplicada a los datos de metabolómica (y otras ómicas), en la que se realiza una estandarización de los datos por *feature* en lugar de por muestra. Por ello, hemos preferido trabajar con los valores de X’ y aplicar la estandarización a nivel de proteína.

DChart

Description automatically generated

Distribución de los valores de la matriz antes y después de aplicar el centrado y escalado a cada proteína. [Enlace](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Proteomics/PESA/WorkingFiles/Plots/SummaryPlots.html) a datos completos.

**Filtro e imputación de missing values**

El número total de proteínas identificadas fue 6213. No obstante, solo 470 proteínas tenían un porcentaje de missing value inferior al 20%, por lo que decidimoas aplicar este filtro e imputar los missing values aplicando KNN. En total se imputó un 3.02% de los datos (6,024 / 199,280).

**Corrección del Batch Effect**

Para comprobar la influencia de diferentes variables en la clusterización de los datos hemos realizado una PCA con la matriz y hemos analizado la correlación entre la proyección de las observaciones sobre las componentes principales y los metadatos [1]. Estos resultados muestran un efecto batch muy acusado, por lo que aplicamos el método ComBat para corregirlo [2].





p-valor del F-test obtenido al construir el modelo lineal:

Antes y después de aplicar ComBat para corregir el efecto batch (columna Cohort).

Graphical user interface, application, Teams

Description automatically generated

Chart, scatter chart

Description automatically generated

[Enlace](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Proteomics/PESA/WorkingFiles/Plots/PCAPlots.html) a todos los PCA.

### Metabolómica

**Generación y preprocesado de los datos**

Se realizó un experimento de metabolómica no dirigida empleando columnas HILIC y C18 en modo positivo y negativo, por lo que tras la detección y alineamiento de las features se obtienen cuatro matrices de intensidades:

Las muestras se inyectaron en el equipo alternando casos y controles e intercalando cada cierto número de muestras un quality control (QC). Esto permite aplicar el método quality control-based robust LOESS signal correction (QC-RLSC) para corregir la desviación de señal de las features en cada muestra [3]. Para ello, hemos empleado la función shiftCor del paquete de R statTarget [4]. Para cada feature, esta corrección usa los QC para ajustar, mediante LOESS, una curva que relaciona el orden de inyección en el equipo y la intensidad medida. Sobre la curva obtenida se realiza una interpolación cúbica por partes que permite obtener la intensidad de referencia asociada a cada momento de inyección (i.e. a cada muestra). Finalmente, se divide la intensidad de la feature en cada muestra entre su intensidad de referencia. De esta manera, si es la matriz de intensidades para las N observaciones y las F features, y es la matriz con las intensidades de referencia obtenidas mediante el método descrito anteriormente, el modelo QC-RLSC no permite obtener una matriz corregida tal que:

Por último, y siguiendo el flujo de normalización aplicado a los datos de proteómica, podemos obtener el logaritmo del ratio, que permite aproximar los datos a una distribución normal:

Chart

Description automatically generated

Distribución de los datos de sucesivas matrices obtenidas a lo largo de preprocesado. Figura en el siguiente [enlace](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metabolomics/PESA/WorkingFiles/Plots/DataDistribution.html).

Chart, scatter chart

Description automatically generated

PCA conjunta de las observaciones y los QC con los datos de HILIC Positivo tras haber aplicado LOESS+log2+CenterScal. Figuras para todas las plataformas en este [enlace](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metabolomics/PESA/WorkingFiles/Plots/PCA_QC.html).

**Filtro e imputación de missing values**

Se eliminaron las features con un coeficiente de variación superior al 30% y más de un 20% de missing values. Asimismo, la imputación se realizó empleando el método KNN.



Número de observaciones y número de features en cada plataforma. En C18P se produjo una caída de la señal en uno de los batches. Dado que trabajaremos con una única matriz de datos de metabolómica, al unificar las cuatro matrices perderemos esas observaciones en todos los casos.

**Corrección del Batch Effect**

Aplicamos el mismo procedimiento descrito en proteómica en el que se analiza la correlación entre la proyección de las observaciones sobre las componentes principales y los metadatos [1].

Estos resultados muestran un efecto batch muy acusado, por lo que aplicamos el método ComBat para corregirlo [2].





p-valor del F-test obtenido al construir el modelo lineal:

Antes y después de aplicar ComBat para corregir el efecto batch (columna Cohort).

Graphical user interface, chart, application, scatter chart

Description automatically generated

Graphical user interface, chart, application, scatter chart

Description automatically generated

Enlace a las [PCA](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metabolomics/PESA/WorkingFiles/Plots/PCA.html) y las distribuciones [antes](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metabolomics/PESA/WorkingFiles/Plots/CohortBatchEffect.html) y [después](S://U_Proteomica/UNIDAD/software/MacrosRafa/data/Metabolomics/PESA_Integromics/Data/Metabolomics/PESA/WorkingFiles/Plots/CohortBatchEffectComBat.html) .

**Anotaciones putativas**

Las anotaciones putativas de las features se realizaron empleando la API de Ceu Mass Mediator 3.0 (<http://ceumass.eps.uspceu.es/mediator/api/v3/batch>) con la siguiente configuración:

**Negative**

{

"metabolites\_type": "all-except-peptides",

"databases": ["all-except-mine"],

"masses\_mode": "mz",

"ion\_mode": "negative",

"adducts": ["M-H", "M-2H", "M+Cl", "M-H-H20", "M-H+HCOONa", "M+Na-2H"],

"tolerance": 10.0,

"tolerance\_mode": "ppm",

"masses": [m1, …, mn]

}

**Positive**

{

"metabolites\_type": "all-except-peptides",

"databases": ["all-except-mine"],

"masses\_mode": "mz",

"ion\_mode": "positive",

"adducts": ["M+H", "M+2H", "M+Na", "M+K", "M+H-H2O", "M+H+HCOONa"],

"tolerance": 10.0,

"tolerance\_mode": "ppm",

"masses": [m1, …, mn]

}

Los resultados de CMM fueron procesados y simplificados usando los cuatro módulos de TurboPutative (Tagger, REname, RowMerger y TPMetrics). La mayoría de las anotaciones putativas que aparecen en las tablas posteriores son las obtenidas en la tabla final filtrada generada por TurboPutative (TPFilter).

# Exploratory Factor Analysis using MOFA

El Análisis de Factores (FA, Factor Analysis) es un método estadístico para la reducción no supervisada de la dimensionalidad que facilita la identificación y descripción de conjuntos de variables altamente correlacionadas. El método permite describir la variabilidad observada en términos de variables latentes o factores (en PCA lo llamaríamos componentes principales) cuyo valor depende de variables que presentan un comportamiento similar en las observaciones. El *loading* de una variable es una medida de la asociación de dicha variable al factor considerado. Si múltiples variables tienen un *loading* elevado para un mismo factor, será un indicador de que estas variables están correlacionadas, algo que podremos comprobar y visualizar mediante un heatmap como veremos a continuación. Aunque FA y PCA presentan semejanzas, en este último el cálculo de las componentes principales está muy influido por las features que presentan una alta variabilidad en las observaciones, aunque su correlación sea muy baja, algo que se logra mitigar en el FA.

El modelo MOFA (Multi-Omics Factor Analysis) se apoya en esta idea y la extiende y generaliza para datos multi-ómicos, integrados por matrices confeatures de distinta naturaleza. Esto permite identificar biomóleculas de diferentes niveles funcionales correlacionados en las observaciones. MOFA aplica un algoritmo iterativo que permite estimar el número óptimo de factores y además aplica una regularización que logra reducir el número de variables con *loading* no nulo en los diferentes factores, lo que facilita en gran medida la interpretación de los datos.

Hemos aplicado MOFA a los datos de proteómica y metabolómica de PESA y AWHS con los siguientes resultados.

## PESA

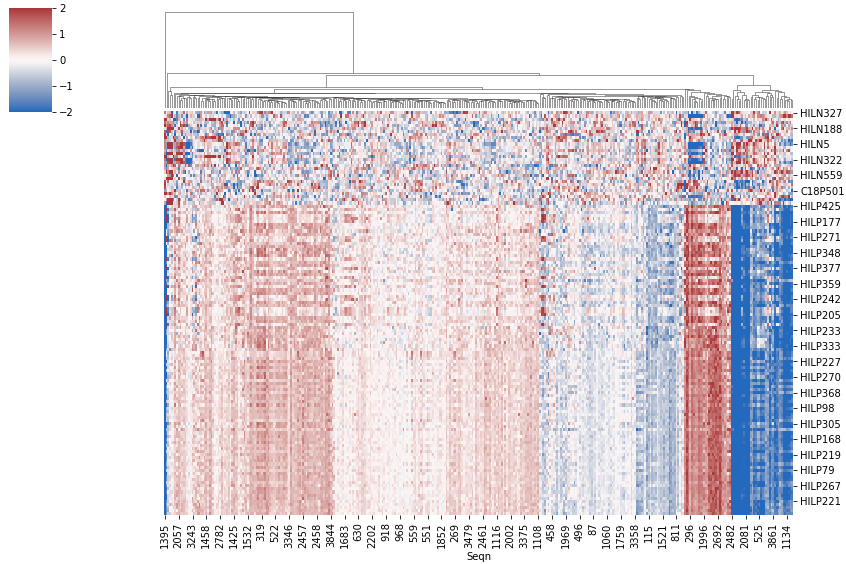
La aplicación de MOFA en PESA permite obtener 11 factores que logra explicar un 50.702% de la variabilidad de los datos en metabolómica y un 10.655% en proteómica. Asimismo, al realizar un análisis de correlación entre los factores y los metadatos de los pacientes, observamos fuertes asociaciones en los factores 2 y 4, y en menor medida en los factores 5, 6 y 8.



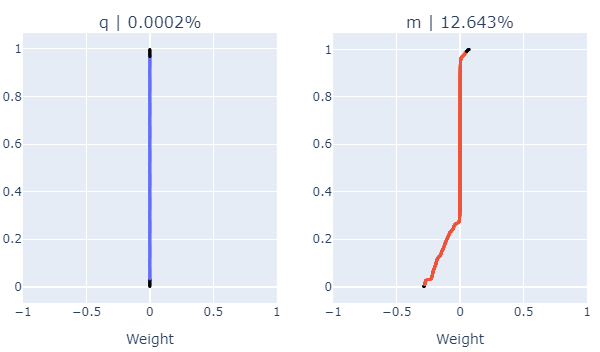
Tabla 1. Análisis de correlación entre los factores y los metadatos de los pacientes. Los valores representados son los p-valores ajustados mediante BH de un F-test aplicado sobre el modelo de regresión lineal: Factor = a\*Metadata + b

Los factores 1 y 3 no aparecen asociados a ninguna de las condiciones medidas en los pacientes aunque resultan muy informativas al explicar una elevada variabilidad de los datos de metabolómica, como vemos a continuación.

**Factor 1**

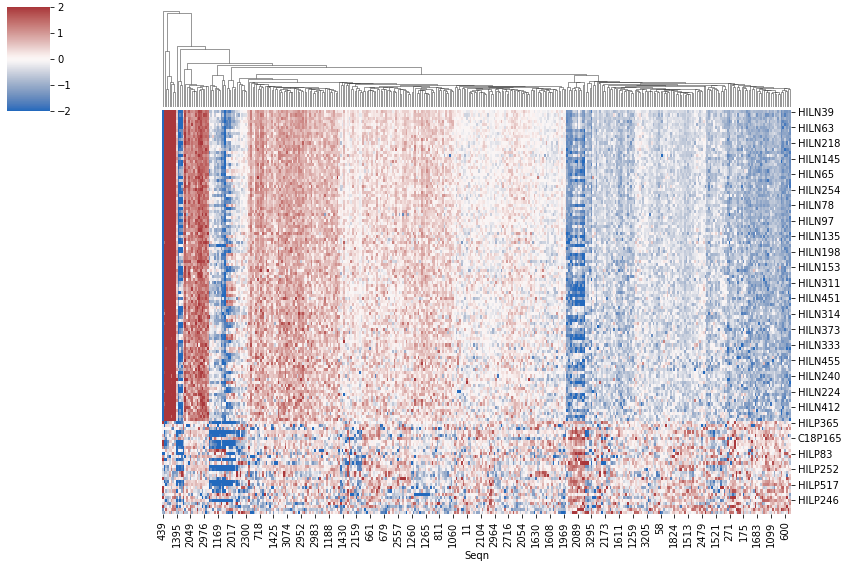


Heatmap 1. Valores de los metabolitos ([30](../Data/Analysis/04-FactorAnalysis/PESA/Tables/Factor1_m_up.tsv), [-100](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor1_m_down.tsv)) y proteínas (0, 0) más relevantes del Factor 1 del en las observaciones. En este caso solo metabolitos.

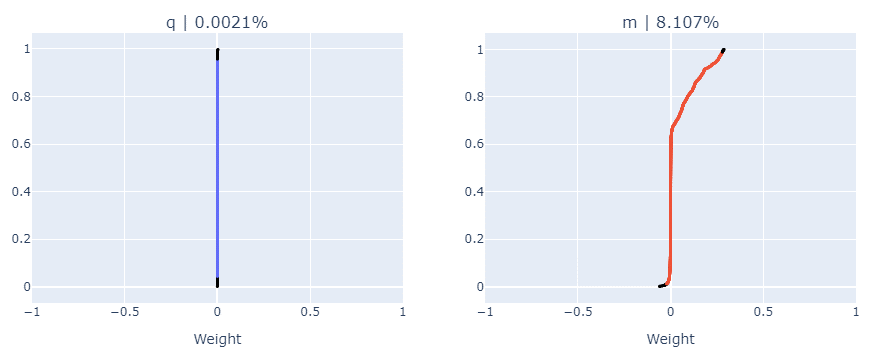


*Loadings* 1. Valores de los *loadings* de los metabolitos en el factor 1.

**Factor 3**



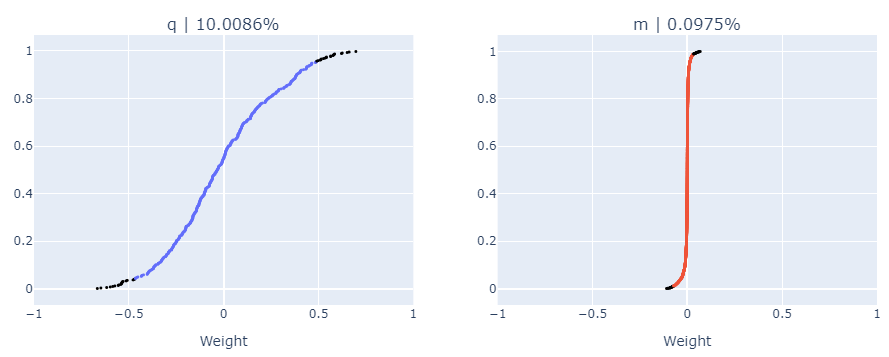
Heatmap 2. Valores de los metabolitos ([100](file:///\\tierra.cnic.es\SC\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor3_m_up.tsv), [-30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor3_m_down.tsv)) y proteínas (0, 0) más relevantes del Factor 3 del en las observaciones. En este caso solo metabolitos.



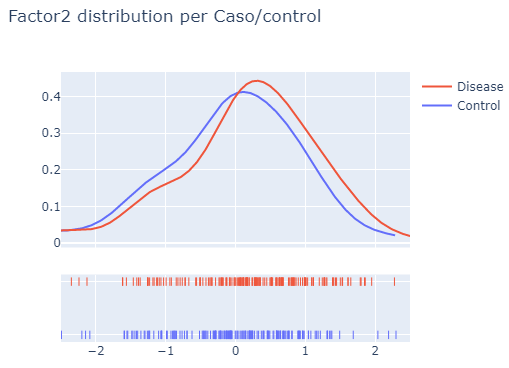
*Loadings* 2. Valores de los *loadings* de los metabolitos en el Factor 3.

**Factor 2**

Como observábamos más arriba, el Factor 2 es el que presenta una mayor asociación a la condición aterosclerótica. Este factor explica una gran parte de la variabilidad de los datos de proteómica y solo una pequeña fracción en metabolómica. Aun así, encontramos algunas features que, aunque con bajo peso, participan en este factor.



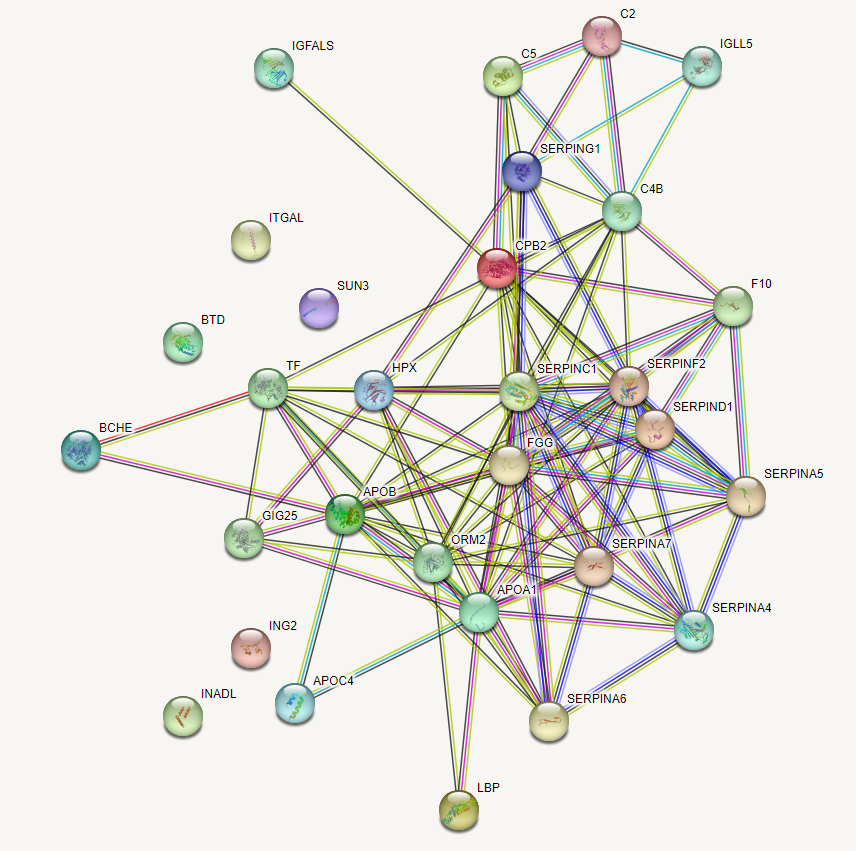
*Loadings* 3. Valores de los *loadings* de los metabolitos en el Factor 2.



Distribución de los valores del Factor 2 distinguiendo entre casos y aterosclerosis subclínica.

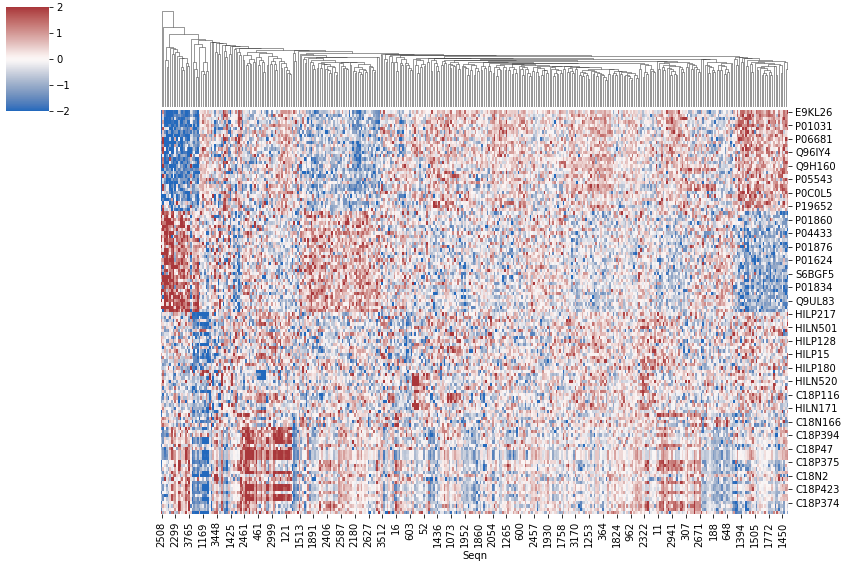
Análisis de enriquecimiento de las proteínas:





Strings 1. Análisis de enriquecimiento con proteínas (30, -30) del Factor 2 mediante test hipergeómetrico usando como background todas las proteínas identificadas:

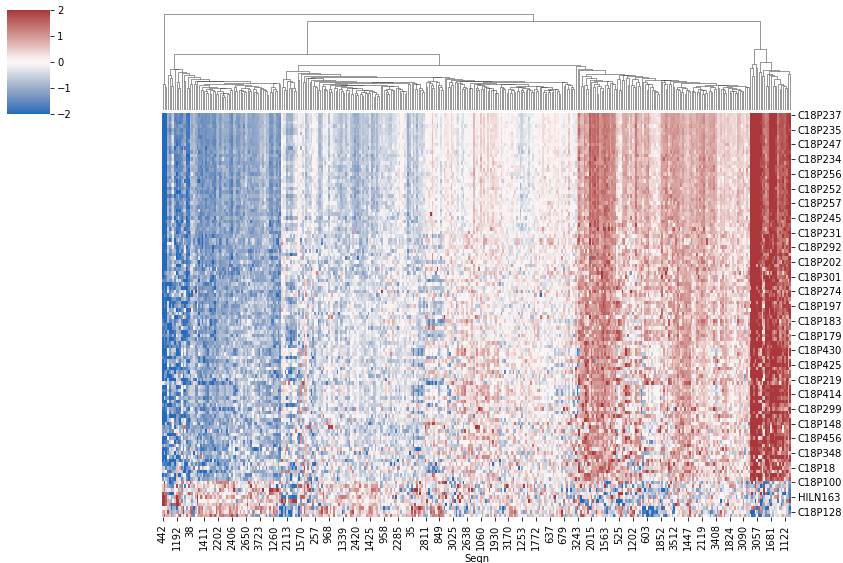
[30 items (human) - STRING interaction network (string-db.org)](https://version-11-5.string-db.org/cgi/network?taskId=bOICM8FcWfl4&sessionId=bwzCEu3MuY3O)



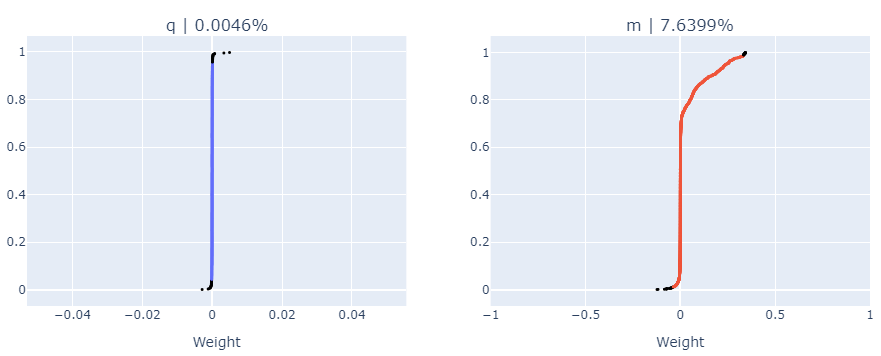
Heatmap 3. Valores de los metabolitos ([30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor2_m_up.tsv), [-30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor2_m_down.tsv)) y proteínas ([30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor2_q_up.tsv), [-30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor2_q_down.tsv)) más relevantes del Factor 3 del en las observaciones. En este caso solo metabolitos.

**Factor 4**

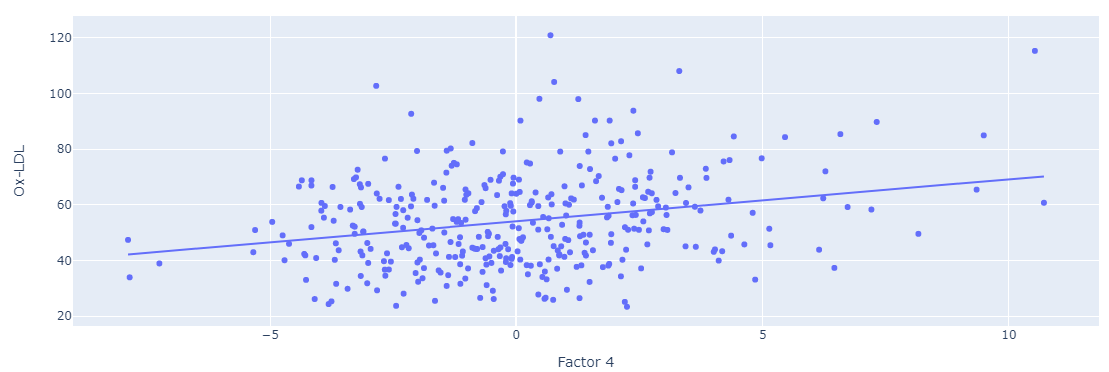
El análisis de correlación muestra asociación entre el cuarto factor y los niveles de colesterol de los pacientes, especialmente LDL y LDL oxidado. Este factor está principalmente ligado a las variables de metabolómica. La features con mayor peso tienen anotados esteroles y otros lípidos oxidados. Por otro lado, el análisis de enriquecimiento de proteínas muestra la participación de apolipoproteínas y otras proteínas implicadas en el metabolismo de lípidos.



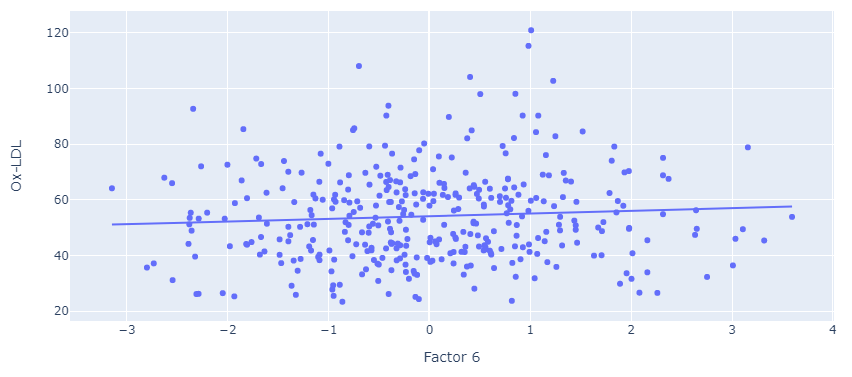
Heatmap 4. Valores de los metabolitos ([100](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor4_m_up.tsv), [-10](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor4_m_down.tsv)) y proteínas (0, 0) más relevantes del Factor 3 del en las observaciones. En este caso solo metabolitos.



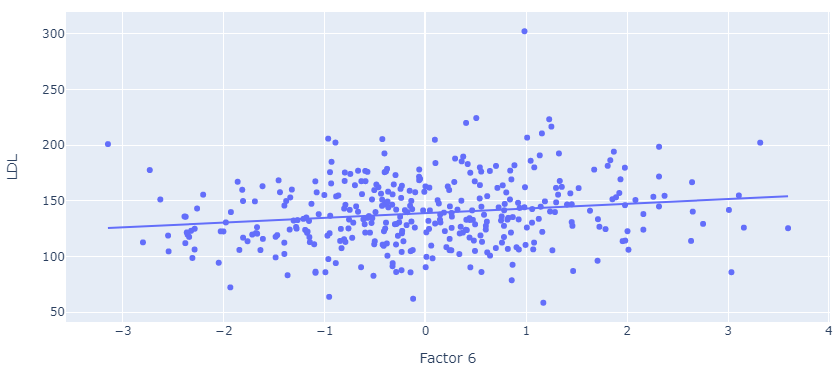
*Loadings* 4. Valores de los *loadings* de los metabolitos en el Factor 4.



Asociación entre el Factor 4 y la LDL-Oxidada.

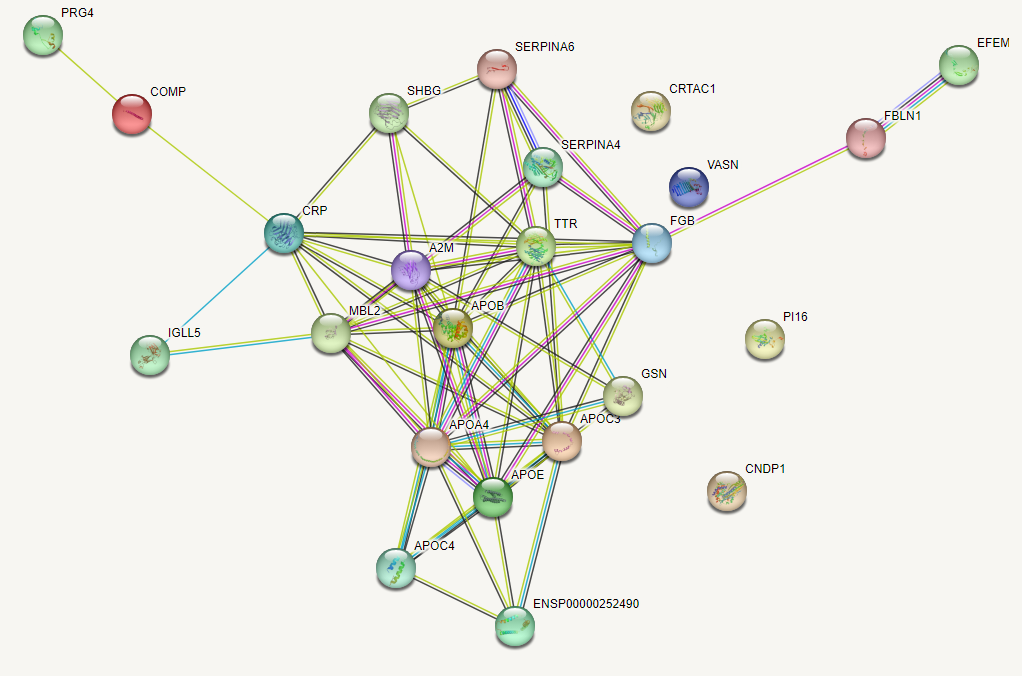


Asociación entre el Factor 6 y la LDL-Oxidada.



Asociación entre el Factor 6 y la LDL.





Strings 2. Análisis de enriquecimiento de proteínas (15, -15) mediante test hipergeómetrico usando como background todas las proteínas identificadas:

[24 items (human) - STRING interaction network (string-db.org)](https://version-11-5.string-db.org/cgi/network?taskId=bEAsRKs8De8c&sessionId=bLjNsxrIiat0)

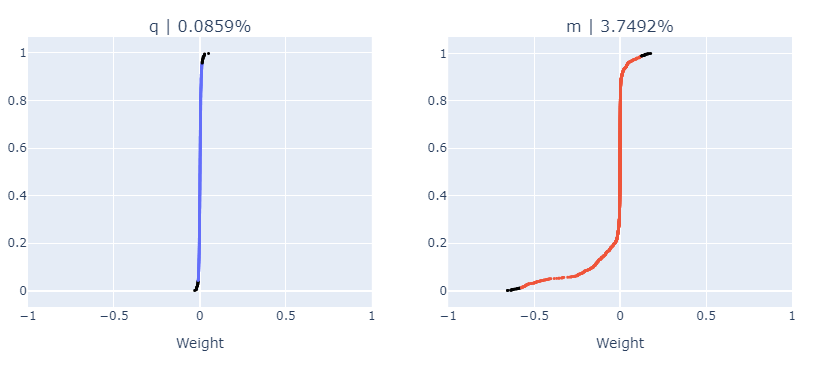
**Factor 5**

La mayor parte de las features asociadas al factor 5 son UNK.

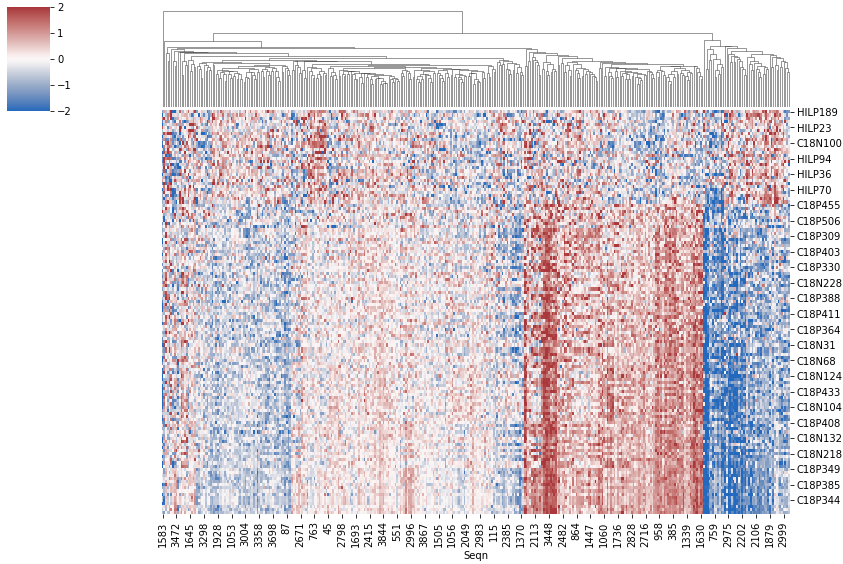
**Factor 6**

Al igual que el Factor 4, el Factor 6 aparece asociado a los niveles de colesterol de las observaciones, especialmente LDL. En este caso no aparece asociación al LDL oxidado. El análisis de enriquecimiento de proteínas sale también parecido aunque con menos proteínas.

En este caso puede ser interesante comparar lípidos con *loading* negativo y positivo.

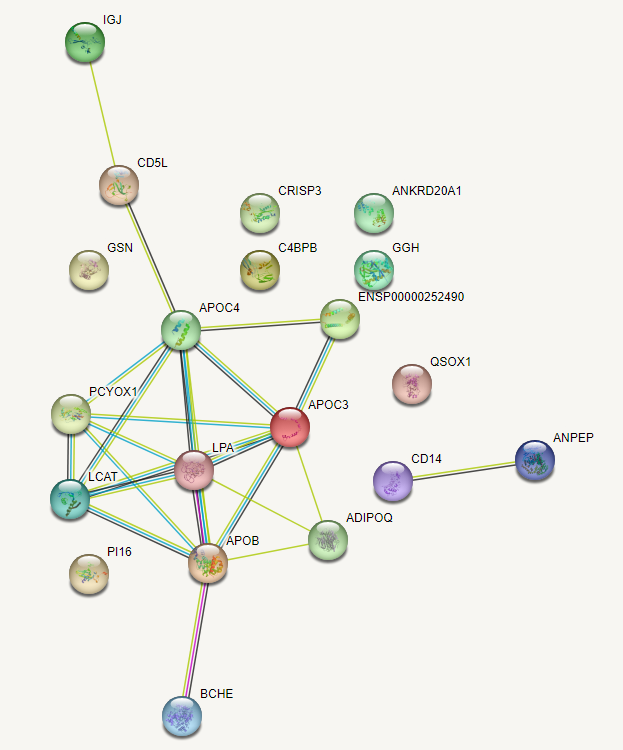


*Loadings* 5. Valores de los *loadings* de los metabolitos en el Factor 4.



Heatmap 5. Valores de los metabolitos ([30](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor6_m_up.tsv), [-100](file:///S:\U_Proteomica\UNIDAD\software\MacrosRafa\data\Metabolomics\PESA_Integromics\Data\Analysis\04-FactorAnalysis\PESA\Tables\Factor6_m_down.tsv)) y proteínas (0, 0) más relevantes del Factor 3 del en las observaciones. En este caso solo metabolitos.





Strings 3. Análisis de enriquecimiento de proteínas (20, -20) mediante test hipergeómetrico usando como background todas las proteínas identificadas:

[20 items (human) - STRING interaction network (string-db.org)](https://version-11-5.string-db.org/cgi/network?taskId=bklrIgExpC46&sessionId=bgiB4msrlwSI)

# Differential Correlation Network Analysis

* El objetivo de esta sección es analizar la influencia de la aterosclerosis subclínica en las interacciones entre proteínas y metabolitos así como su posible asociación a los diferentes metadatos de los pacientes.
* Ello lo haremos mediante la construcción de un grafo donde los nodos representan las proteínas y las features detectadas en los experimentos de proteómica y metabolómica, respectivamente. Asimismo, dos nodos aparecerán conectados si su correlación se ve significativamente alterada en los casos (aterosclerosis subclínica) respecto a los controles.
* Explicar cómo se obtiene
* Para determinar la significatividad estadística de la diferencia de correlaciones, , planteamos un contraste de hipótesis en el que la hipótesis nula es la no influencia de la condición caso/control sobre la correlación diferencial. Así, la distribución de bajo la hipótesis nula se estima mediante Monte Carlo permutation test. Para ello, calculamos los diferentes valores de tras realizar múltiples permutaciones sobre los datos originales (para eliminar la influencia caso/control). Así, si i y j son nodos del grafo descrito anteriormente cuya correlación diferencial entre caso y control es , los nodos i y j aparecerán conectados si y solo si .

## AWHS

### Metabolómica

Es un experimento de la metabolómica semidirigida, que se centró en la detección y cuantificación de un grupo amplio de lípidos. Para la cuantificación, se añadió sobre cada una de las muestras un conjunto de estándares internos, cada uno de los cuales permite estimar la concentración de los lípidos correspondientes a su misma clase lipídica.

La cuantificación se realizó a nivel de precursor y de fragmentación, siendo esta última la más fiable al evitar posibles interferencias con otros lípidos distintos de igual m/z. Será, por tanto, con esta con la que hagamos los análisis posteriores (recomendación de Alessia).

Pensemos en la tabla primaria de los datos como una matriz FxN (F=número de features, N=número de observaciones), que asigna a cada feature y observación una medida de la intensidad de la señal detectada en los equipos. Tendríamos una matriz para las señales obtenidas al nivel del precursor y otra para las señales obtenidas al nivel de la fragmentación. Aunque estamos hablando de feature, al tratarse de un experimento de metabolómica semidirigida, cada una de estas features está perfectamente identificada y asignada a un lípido concreto.

Los valores de intensidad de cada una de las features en una muestra m se transformaron de la siguiente forma:

Donde f, r y m denotan la feature, su clase lipídica de referencia y la muestra, respectivamente. Veamos que la concentración de la feature se obtiene dividiendo su medida de intensidad entre la del estándar interno y multiplicando por su concentración. Al igual que en proteómica, se trabajará con el logaritmo del ratio. Al aplicar un centrado y escalado a los valores de la feature (en escala logarítmica), el efecto de la concentración desaparece, por lo que estaremos trabajando con el logaritmo del ratio estandarizado al igual que hacemos en Proteómica, lo que facilita el proceso de integración multi-ómica.

En este experimento debemos considerar la existencia de tres tipos de missing values, gestionados de manera distinta:

- Casos en los que no se detectó el compuesto con el equipo (pero sí el estándar interno). Lo que se hizo fue imputar la concentración asignando 0.00001. Al aplicar log es -5.

- Casos en los que no se produjo la fragmentación en ningún caso (None). Los eliminamos de la tabla MS2.

- Casos en los que no se ha detectado el estándar interno (NaN). Imputamos por KNN.

Análisis Exploratorio de los Datos:

* Distribución de la concentración a nivel de MS1 y MS2, distinguiendo ionización positiva y negativa. Asimismo, distinguimos por cohorte y condición control/enfermo.
* Scatter plot en el que representamos la fracción de missing values imputados con -5 para cada feature en control vs enfermo.
* Distribución de los log-ratios antes y después de estandarizar, distinguiendo por cohorte y caso.
* Reducción de la dimensionalidad mediante PCA y UMAP, distinguiendo por cohorte y caso.

# Referencias

[1] L. Gatto *et al.*, “Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments,” Jul. 2022, doi: 10.1038/s41592-023-01785-3.

[2] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/biostatistics/kxj037.

[3] W. B. Dunn *et al.*, “Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry,” *Nat Protoc*, vol. 6, no. 7, pp. 1060–1083, Jun. 2011, doi: 10.1038/nprot.2011.335.

[4] H. Luan, F. Ji, Y. Chen, and Z. Cai, “statTarget: A streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data,” *Anal Chim Acta*, vol. 1036, pp. 66–72, Dec. 2018, doi: 10.1016/j.aca.2018.08.002.