PEC 2 - Diseño y Análisis de Experimentos

Celia Martínez Saz

Tabla de contenido

Resumen 1

Objetivos 1

Métodos 2

Resultados 4

Discusión 5

Conclusiones 7

Referencias 8

Resumen

La caquexia es un síndrome metabólico multifactorial con que caracteriza a la pérdida muscular con o sin pérdida de grasa. Dado que varios metabolitos del catabolismo muscular son excretados por la orina y por su toma de muestras poco invasiva, este trabajo se ha centrado en el estudio de metabolitos presentes en la orina mediante RMN unidimensional en 77 muestras de orina de pacientes con caquexia (n=47) y personas sanas (n=30). Para llevar a cabo este estudio se empleó la herramienta Bioconductor y se trabajó con diversos tipos de análisis multivariante. Los metabolitos resulta

Objetivos

La caquexia es un síndrome metabólico que, en ocasiones, se asocia a enfermedades subyacentes como el cáncer y produce con la pérdida progresiva de masa muscular. Los métodos de evaluación actuales son técnicas costosas o invasivas como la tomografía computarizada. Por este motivo existe la necesidad de búsqueda e identificación de biomarcadores para facilitar su detección.

Por ello, el objetivo principal de este proyecto es la identificación de biomarcadores urinarios que faciliten el diagnóstico de la caquexia. Para ello, se ha evaluado el perfil de 63 metabolitos urinarios diferentes mediante RMN-1H como posible herramienta de detección de la caquexia. La elección de muestras de orina se debe a que muchos productos del catabolismo muscular se excretan por esta vía, y su obtención es menos invasiva que la extracción sanguínea o la biopsia muscular, necesarias para el estudio de biomarcadores sanguíneos o intramusculares. Por tanto, el estudio metabolómico en orina podría representar una alternativa menos invasiva y más accesible a los métodos actuales.

Concretamente, el estudio se centra en detectar diferencias y comparar los perfiles metabólicos de pacientes con caquexia y sujetos sanos mediante herramientas bioinformáticas como Bioconductor, junto con diversos análisis estadísticos. para dar una interpretación biológica de estas diferencias con el propósito mejorar el diagnóstico

Materiales y métodos

Naturaleza de los datos

Los datos pertenecen a un dataset llamado human_caquexia.csv del repositorio GitHub proporcionado para este ejercicio. Se trata de un archivo en el que tras adquirir los espectros de RMN unidimensional de las muestras de orina de los sujetos en estudio (n=77) tanto sanos (n=30) como con caquexia (n=47), se detectaron dichos metabolitos y se midió la concentración de cada uno de los 63 metabolitos en estudio. Por tanto, estos datos nos proporcionan información acerca de los sujetos (ID y pérdida muscular) y las concentraciones de los metabolitos urinarios. Por tanto, nos encontramos ante un problema de metabolómica. Cabe destacar del dataset elegido que las muestras no están emparejadas, hay dos grupos en las muestras (control/caquexia), todos los valores son numéricos y no hay *missing values*.

Metodología empleada

Los datos han sido analizados empleando el software RStudio y herramientas y librerías de Bioconductor, un proyecto de código abierto para el análisis de datos ómicos. Para el diseño de matrices y la creación del SummarizedExperiment he adaptado y consultado diferentes códigos (1,2,3,4), así como los recursos del curso. Todo el código empleado se encuentra en el archivo “RMarkdown-creacion-SummarizedExperiment--y-metadatos” en GitHub debidamente comentado.

Cabe destacar que la clase de Bioconductor SummarizedExperiment se emplea para almacenar y gestionar resultados experimentales de forma estructurada y coordinada. Este tipo de clase contiene diferentes elementos:

- Assays (datos experimentales): son las matrices con los resultados numéricos de los experimentos, en nuestro caso, las concentraciones de los diferentes metabolitos en orina. También, puede manejar varios resultados experimentales a la vez, siempre que la matriz tenga las mismas dimensiones. Por ejemplo, si tuviéramos datos de los mismos metabolitos en sangre o intramusculares, podríamos almacenarlos también.
- Encontramos dos tipos de metadatos:

- Metadatos que describen las características y se almacenan en las filas (rowData), en nuestro caso el ID de paciente y la pérdida muscular,
- Metadatos que describen las muestras (colData) y se almacenan en las columnas. Un ejemplo podría ser los metabolitos urinarios del estudio. (3,5)

Este tipo de objeto es similar al ExpressionSet pero la principal diferencia es que es más flexible en cuanto a la información de las filas. SummarizedExperiment permite tanto GRanges (para rangos genómicos, coordenadas de genes entre otros) como Dataframes, siendo adecuado para una gran variedad de experimentos; desde datos genómicos (RNA-seq, ChIP-Seq) hasta no genómicos (metabolómica, proteómica, lipidómica...).

Por otro lado, ExpressioSet se emplea generalmente para datos procedentes de expresión génica basados en matrices, donde las filas representan genes y las columnas muestras. La diferencia es que los metadatos de muestras y genes se almacenan como objetos separados: phenoData (metadatos de muestras) y featureData (metadatos de genes). Además, los nombres de las columnas del objeto que contiene las expresiones (assayData) deben coincidir con los nombres de las filas de phenoData.

No obstante, cada uno de ellos tiene sus ventajas; los ExpressionSets tienen una estructura que garantiza su consistencia. Mientras SummarizedExperiment es más flexible, pero también permite un almacenamiento más estructurado y organizado para datos y metadatos, así como la posibilidad de almacenar más de un experimento. (3,4,5)

Herramientas estadísticas y bioinformáticas utilizadas

Para llevar a cabo la creación de SummarizedExperiment se ha empleado el software R y Bioconductor. Para el análisis estadístico, primero se ha llevado a cabo un preprocesamiento de datos normalizándolos mediante un logaritmo natural, dado que las concentraciones de metabolitos urinarios suelen no seguir una distribución normal (8). A continuación, se ha llevado a cabo un análisis de componentes principales, basándonos ene el código proporcionado por el material de estudio (https://aspteaching.github.io/AMVCasos/#presentaci%C3%B3n) (11) y otros recursos (https://anoteweb2.bio.di.uminho.pt/metabolomicspackage/cachexia/cachexia_noproc.html) (14). Este código se encuentra debidamente comentado en el documento “Análisis exploratorio Human Caquexia” en Git Hub.

Procedimiento general de análisis

El análisis se ha realizado siguiendo el siguiente procedimiento:

1. Cargar los datos a R
2. Creación de SummarizedExperiment con Bioconductor
3. Preprocesamiento de datos: normalización
4. Calculo de componentes principales
5. Pruebas T para cada metabolito con ajuste de Bonferroni

Resultados

Tras realizarse un preprocesamiento de los datos, transformando las concentraciones de los metabolitos urinarios a escala logarítmica para normalizar estos datos, se ha llevado a cabo una serie de análisis estadísticos del dataset Human_cachexia.

Análisis de componentes principales

En primer lugar, se ha realizado un cálculo de los componentes principales (PCA) y finalmente se ha representado el PC1 frente al PC2 (Figura 1). Tras realizar el análisis de componentes realizado (PCA) y representar el PC1 frente al PC2 observamos dos grupos más o menos diferenciados; por un lado, encontramos las muestras correspondientes de pacientes sanos en azul claro, mientras que los pacientes con caquexia se encuentran representados por triángulos morados. En este gráfico podemos ver que los sujetos control se encuentran mayormente en la parte izquierda de la gráfica (valores negativos), mientras que las muestras de sujetos control predominan en la parte derecha (valores positivos). No obstante, podemos ver que el grupo de los sujetos control presenta una mayor variabilidad al encontrarse más disperso a lo largo de la gráfica. Finalmente, observamos algunos valores que podrían ser outliers; algunos triángulos morados en la parte izquierda, o por el contrario puntos que representan a sujetos control en la parte derecha. No obstante, esto tendríamos que comprobarlo.

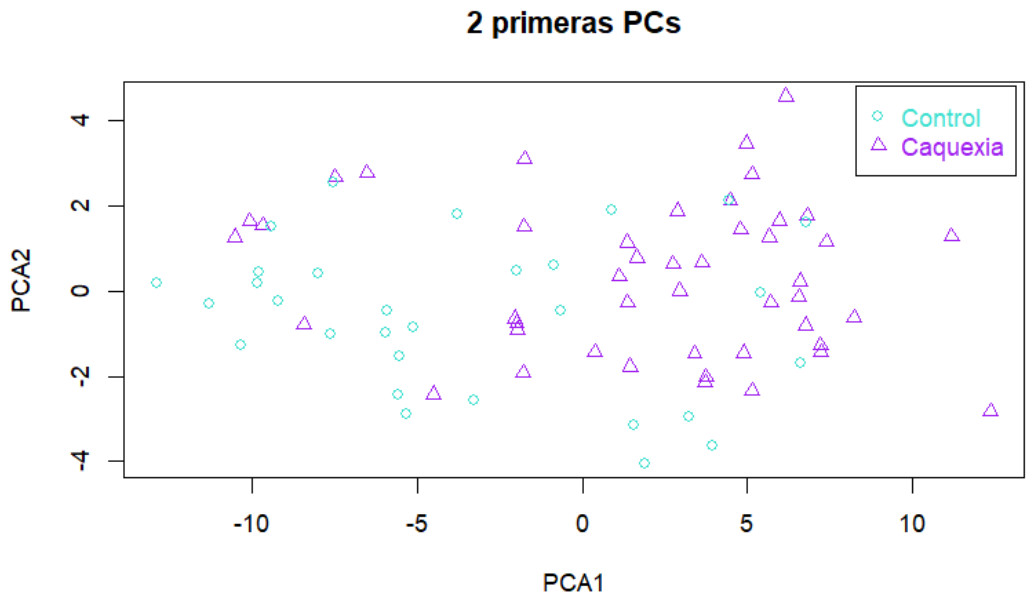

Figura 1. Representación del PCA1 frente al PCA 2.

Finalmente, para determinar los metabolitos que contribuyen más a cada uno de los componentes principales se ha extraído la matriz "rotation" que proporciona los loadings de los componentes principales y siendo los siguientes:

Tabla 1. Metabolitos que influyen más en PCA1

Metabolito	Cis-aconitato	Succinato	Histidina	Creatina	Glutamina
Valor	0.1696246	0.1669296	0.1574637	0.1573320	0.1573131

Tabla 2. Metabolitos que influyen más en PCA1

Metabolito	Acetato	2-Oxoglutarato	Sucrosa	pi-Metilhistidina	Succinato
Valor	0.1696246	0.1669296	0.1574637	0.1573320	0.1573131

Vemos que en el PC1 el metabolito que más influye es el cis-aconitato, mientras que en el PC2 el acetato. Esto significa que, por un lado, el cis-aconitato tiene gran influencia el PC1, mientras que el acetato en el PC2. Destacando que en ambos componentes principales encontramos el succinato, por lo que quizá sea un metabolito clave.

T test

Se llevo a cabo un T-test de comparación de medias entre grupos con ajuste de Bonferroni para identificar los metabolitos que muestran diferencias significativas en sus concentraciones detectadas por RMN-1H entre el grupo de pacientes control (n=30) y el grupo de pacientes con caquexia (n=47). A continuación, se muestran los metabolitos con diferencias significativas (Tabla 3):

Tabla 3: Metabolitos con p-valor < 0.05

Metabolito	**p-valor**	**Metabolito**	**p-valor**
X1.6.Anhydro.beta.D.glucose	0.020045455	Alanine	0.0104297220
X1.Methylnicotinamide	0.076371469	Tryptophan	0.0130947857
X2.Aminobutyrate	0.001764377	Dimethylamine	0.0131399570
X2.Hydroxyisobutyrate	0.003202246	Methylamine	0.0140072454
X2.Oxoglutarate	0.042242446	Betaine	0.0167618971
X3.Aminoisobutyrate	0.153780180	Formate	0.0184376672
Glucose	0.0001615225	Creatinine	0.0269378612
Adipate	0.0006392746	Sucrose	0.0304350955
Quinolinate	0.0007747771	Lactate	0.0421284457
Leucine	0.0011157180		
Valine	0.0016829718		
myo.Inositol	0.0017695007		
X3.Hydroxyisovalerate	0.0027080453		
N.N.Dimethylglycine	0.0027912162		
X3.Hydroxybutyrate	0.0037425317		

Como vemos, los aminoácidos con diferencias más signficativas se corresponden a la Glucosa, adipato, quinolinato, leucina, valina, mioinositol entre otros.

Discusión

Algunas de las limitaciones de este estudio es que hubiera estado interesante poder evaluar como en otros estudios la tasa de cambio muscular (% de pérdida o ganancia a lo largo del tiempo) para poderla asociar al estudio metabólico. Asimismo, hubiera sido interesante poder analizar no solo metabolitos urinarios sino también musculares y sanguíneos.

Por otro lado, siendo crítica con el trabajo elaborado en este proyecto, no he llevado a cabo un buen análisis exploratorio del conjunto de datos. Dado que en la bibliografía consultada se llevaba a cabo otro tipo de análisis de datos similares que quizás hubiera sido más apropiado y hubiera alcanzado mejor los objetivos. Por ejemplo como en la bibliografía consultada podría haber aplicado técnicas de aprendizaje automático para identificar patrones en los metabolitos.

No obstante, en la bibliografía consultada las estadísticas bivariadas identificaron metabolitos relacionados con la pérdida muscular, incluyendo constituyentes y metabolitos musculares (creatina, creatinina, 3-OH-isovalerato), aminoácidos (Leu, Ile, Val, Ala, Thr, Tyr, Gln, Ser) y metabolitos intermediarios. Atendiendo a los resultados de nuestro T test, los metabolitos que presentaban un p-valor más bajo fueron la glucosa, metabolito clave en el metabolismo energético, el adipato asociado a la descomposición muscular y el quinolinato, relacionado con el estrés celular. Por otro lado, se han detectado diferencias en las concentraciones entre pacientes sanos y caquéxicos en aminoácidos como la leucina y la valina, lo cual concuerda con los estudios consultados.

Atendiendo al PCA, querría destacar que en ambos componentes estudiados uno de los metabolitos que más influencia tenía era el succinato, este metabolito al intervenir en el ciclo de Krebs, es clave para la generación de energía lo que podría explicar su influencia en ambos PCAs. Finalmente, el PCA1 parece estar más relacionado con los pacientes de caquexia, siendo el cis-aconitato el metabolito más influyente, lo que indica que diferencias en las concentraciones de este metabolito aumentan la separación en los grupos control y caquexia. Este metabolito también interviene en el ciclo de Krebs, luego una baja concentración conllevará a alteraciones metabólicas, sobre todo en el músculo. Mientras que en el PCA2 parece estar más asociado con pacientes control, siendo el acetato el que más influye. Los valores negativos en PCA2 podrían indicar un buen funcionamiento metabólico, en el que el acetato está a concentraciones adecuadas y su producción no se ve alterada.

Conclusiones

En este estudio se ha intentado analizar las diferencias metabolómicas de pacientes con caquexia e individuos control a partir de muestras de orinas

analizadas con RMN-1. Se ha llevado a cabo un PCA, así como un T -test en el que se han encontrado algunos metabolitos relevantes que podrían ser futuros biomarcadores. No obstante, este análisis ha sido bastante incompleto y se podría haber completado con otro tipo de análisis como clustering entre otros. Así mismo

Referencias

1. https://bioconductor.org/help/course-materials/2015/Uruguay2015/V2-WorkingWithData.html#make-a-summarizedexperiment-object
2. https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html#introduction
3. https://www.bioconductor.org/packages/devel/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html
4. Sanchez-Pla A. Bioconductor classes for working with microarrays or similar data. 2024
5. https://www.sthda.com/english/wiki/expressionset-and-summarizedexperiment
6. https://aspteaching.github.io/AMVCasos/#presentaci%C3%B3n
7. https://www.bioconductor.org/packages/release/bioc/vignettes/POMA/inst/doc/POMA-workflow.html
8. R. Eisner, C. Stretch, T. Eastman, J. Xia, D. Hau, S. Damaraju, R. Greiner, D. S. Wishart, and V. E. Baracos. Learning to predict cancer-associated skeletal muscle wasting from 1h-nmr profiles of urinary metabolites. Metabolomics, 7:25–34, 2010.
9. Yang QJ, Zhao JR, Hao J, Li B, Huo Y, Han YL, Wan LL, Li J, Huang J, Lu J, Yang GJ, Guo C. Serum and urine metabolomics study reveals a distinct diagnostic model for cancer cachexia. J Cachexia Sarcopenia Muscle. 2018 Feb;9(1):71-85. doi: 10.1002/jcsm.12246. Epub 2017 Nov 19. PMID: 29152916; PMCID: PMC5803608.
10. Cao Z, Zhao K, Jose I, Hoogenraad NJ, Osellame LD. Biomarkers for Cancer Cachexia: A Mini Review. Int J Mol Sci. 2021 Apr 26;22(9):4501. doi: 10.3390/ijms22094501. PMID: 33925872; PMCID: PMC8123431.
11. https://aspteaching.github.io/AMVCasos/#presentaci%C3%B3n
12. drr.io/github/SciDoPhenIA/phenomis/man/writing.html
13. https://docs.github.com/es/enterprise-server@3.12/repositories/working-with-files/managing-files/adding-a-file-to-a-repository
14. https://anoteweb2.bio.di.uminho.pt/metabolomicspackage/cachexia/cachexia_noproc.html
15. https://rpubs.com/cristina_gil/pca
16. https://github.com/Celchy/Martinez_Saz_Celia_PEC1