Jimenez-Ferandez-Pedro-PEC1

Pedro Jiménez Fernández

2025-03-21

Índice

Abstract

La caquexia es un síndrome metabólico que se caracteriza por la pérdida significativa de masa muscular que normalmente está asociado al cáncer u otras enfermedades crónicas. Este estudio analiza un conjunto de datos metabolómicos con el fin de poder identificar diferencias en los perfiles metabólicos entre pacientes que presentan caquexia y controles sanos. Mediante el empleo de R y RStudio se construyó un objeto de la clase "SummarizedExperiment" con 63 metabolitos medidos en 77 individuos. Se efectuó un análisis exploratorio de los datos (EDA), seguido de una transformación logarítmica con el fin de reducir la dominancia de metabolitos con la varinanza alta. Después se efectuó un análisis de componentes principales (PCA), cuyo resultado reveló que la primera componente (PC1) explicaba el 58,5% de la variabilidad total. La PC1 hacía una diferenciación parcial entre el grupo de pacientes y el grupo control, siendo la creatinina el principal metabolito contribuyente. Los resultados obtenidos sugieren un perfil metabólico característico en los individuos caquéxicos, que se alinean con los procesos de degradación muscular descritos en la literatura científica (Fearon et al., 2011).

Objetivos

Este estudio tiene como objetivo principal la comparación de los perfiles metabolómicos entre individuos sanos pertenecientes a un grupo control y pacientes con caquexia con el objetivo de identificar patrones metabólicos asociados a estos últimos. Los objetivos específicos son los siguientes:

- La correcta importación y procesado de los datos metabolómicos.
- La construcción de un objeto de tipo "SummarizedExperiment" para el posterior análisis de los datos.
- La realización de un análisis exploratorio de los datos con el fin de comprender las características de los mismos.
- La realización de un análisis de componentes principales para conseguir una reducción de la dimensionalidad y detectar algún patrón metabólico asociado con la caquexia.

Métodos

El estudio se ha llevado a cabo empleando el lenguaje de programación R, versión 4.4.3 dentro del IDE RStudio. Se ha utilizado RMarkdown para la creación de un informe dinámico que contenga el código empleado y permita reproducir los resultados. Para realizar el análisis se emplearon librerías presentes de base en la versión de R 4.4.3, así como librerías adicionales, entre las que se encuentran curl, BiocManager (BioConductor) y SummarizedExperiment, para la descarga de datos, instalación del paquete SummarizedExperiment y creación del objeto de contención de los datos respectivamente.

Primero se descargaron los datos desde un repositorio de GitHub proporcionado por el profesor Alex Sánchez Pla. A continuación se creó el objeto de la clase SummarizedExperiment para contener la matriz de datos de los diferentes metabolitos en los diferentes pacientes junto a sus metadatos, como por ejemplo el grupo al que pertenecían. Posteriormente se realizó un análisis exploratorio inicial de los datos para comprobar las características de los mismos y buscar diferencias entre el grupo de control y los pacientes. Finalmente, se realizó un análisis de componentes principales con el objetivo de reducir la dimensionalidad de los datos y detectar la separación entre ambos grupos en base a sus perfiles metabólicos.

Resultados

Importación de los datos

En primer lugar se descargan los datos. Una vez descargados los datos en el repositorio, se importan para su posterior análisis.

Creación del objeto SummarizedExperiment

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(4): title grupos fuente fecha
## assays(1): metabolites
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(1): metabolite_name
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): patient_id group
```

Tal y como se puede observar, la variable "cachexia_se" contiene 63 variables (metabolitos) y 77 observaciones (pacientes/individuos sanos) distribuídas en filas y columnas respectivamente.

Análisis exploratorio de los datos y PCA (principal component analysis)

Análisis exploratorio de datos (EDA)

A continuación se hace un análisis exploratorio de los datos contenidos en el objeto "cachexia_se".

```
##
## cachexic  control
##       47       30
```

Se puede observar que en total hay 47 individuos con caquexia y 30 individuos pertenecientes al grupo control.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

Aunque en el archivo "description.md" adjunto a los datos se especifica que no existen valores faltantes, se ha realizado una prueba para verificarlo. Tal y como se indica hay ausencia total de valores faltantes o "missing values". A continuación se muestran las estadísticas básicas.

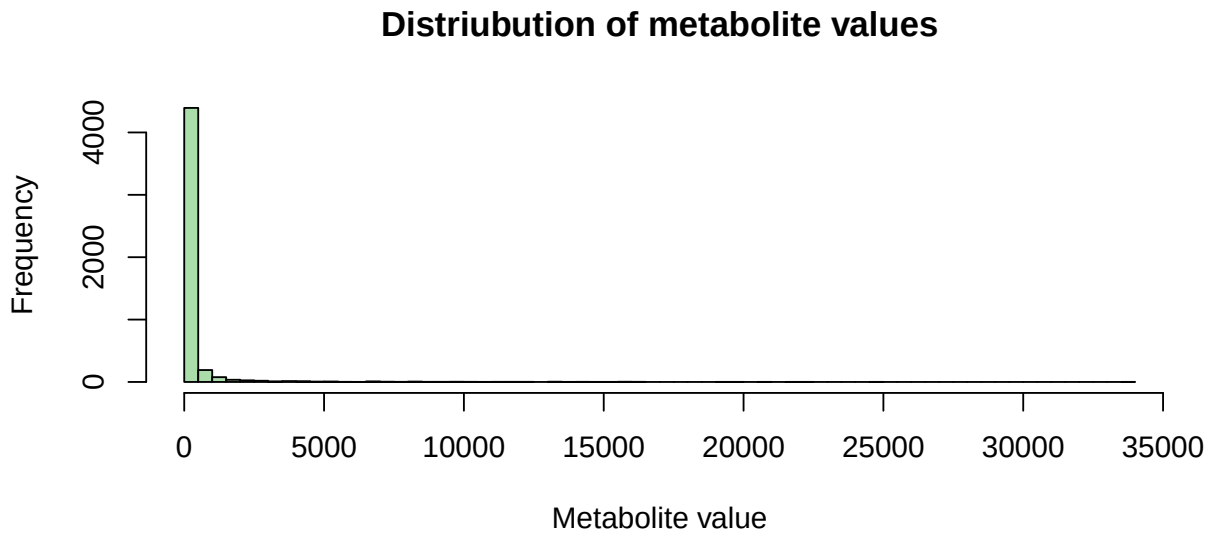
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2696    9296   21563   21885   31266   77965
```

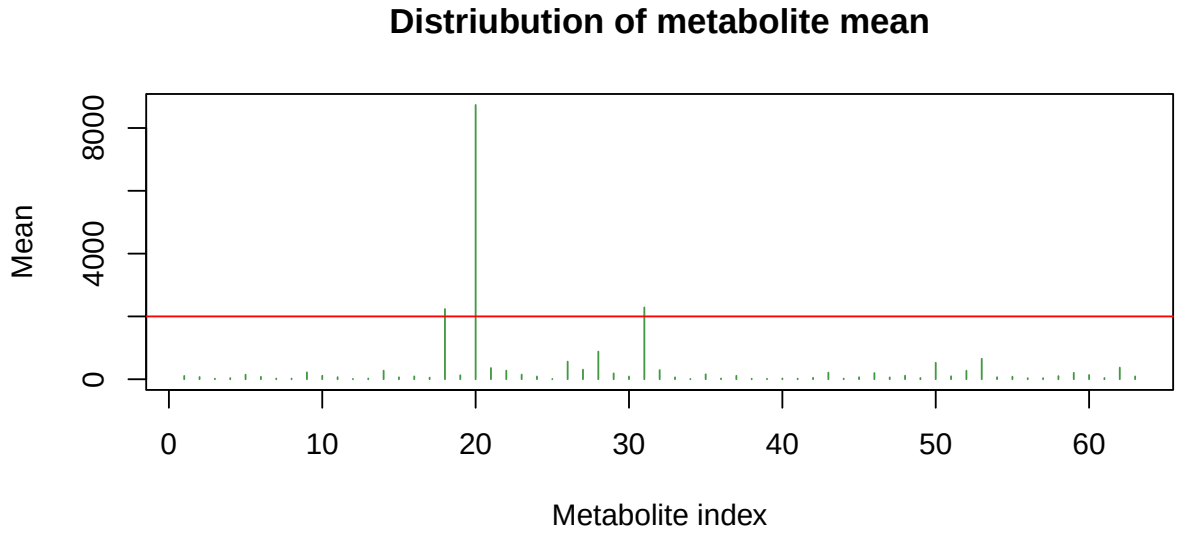
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    649.9   2974.3   6905.9  26747.8  16003.2 672515.8
```

```r
# a continuación, se hace un filtrado de los metabolitos que no presentan
# varianza
variance <- rowSds(matrix)
filtered_matrix <- matrix[variance > 0, ]
nrow(filtered_matrix)
```

```
## [1] 63
```

Los 63 metabolitos presentan varianza. A continuación se representa un histograma global para visualizar la distribución de los valores de los diferentes metabolitos.

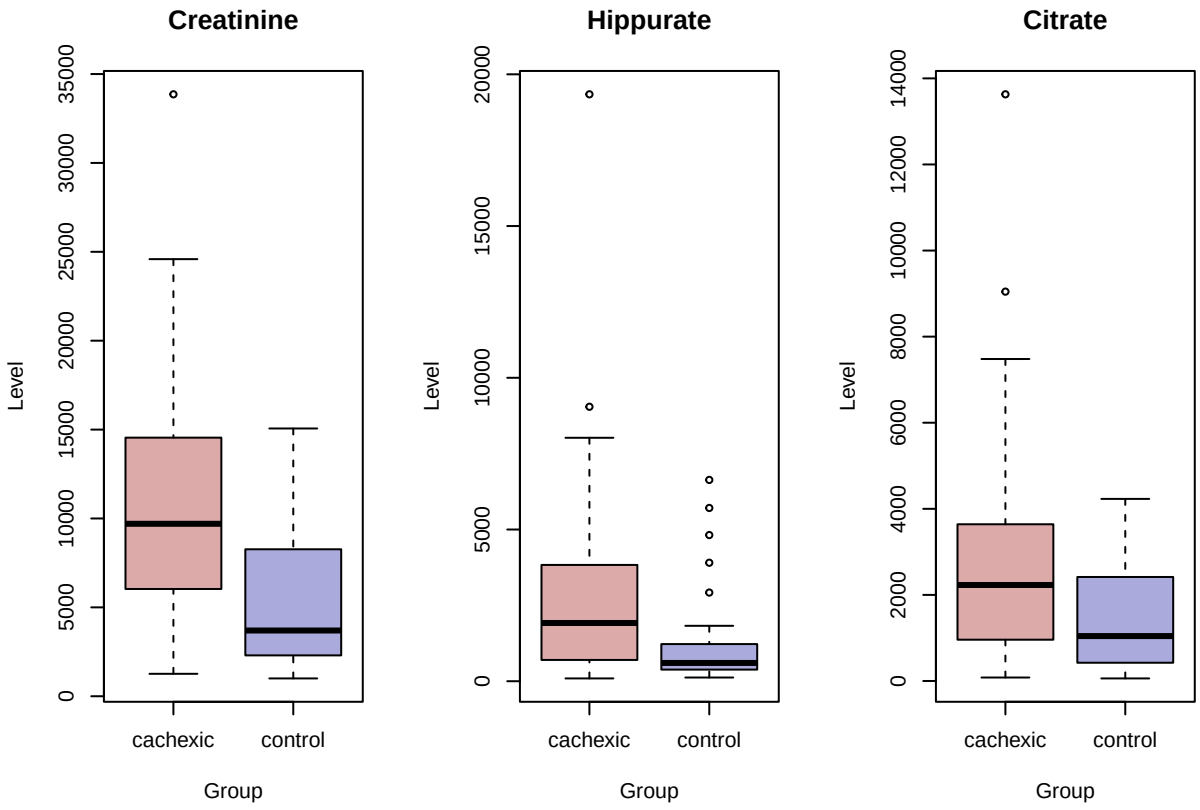




Al observar la distribución global de todos los valores de la matriz, puede observarse que hay un gran número de valores muy cercanos a 0 y después una cola bastante larga que va hasta valores superiores a 30,000. Algunos metabolitos presentan valores muy elevados, lo cual indica la presencia de outliers, por otra

parte, las diferencias en la magnitud de los valores de los metablitos indican que las varianzas no son similares, estos 2 factores contribuyen a que la capacidad de predicción del análisis de componentes principales se vea comprometida. Por otra parte, puede observarse que hay 3 metabolitos cuya media está significativamente por encima de la de los demás metabolitos.

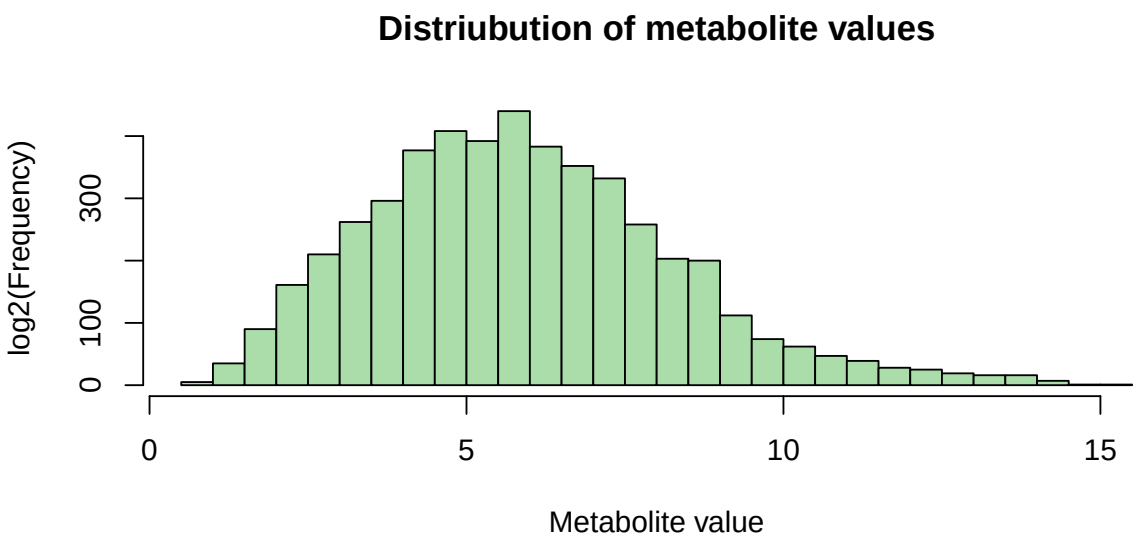
```
##                              mean             metabolite
## Creatinine               8733.9718             Creatinine
## Hippurate                2286.8377              Hippurate
## Citrate                  2235.3460                Citrate
## Glycine                   880.7174                Glycine
## Trimethylamine.N.oxide    652.1569 Trimethylamine.N.oxide
## Glucose                   559.8445                Glucose
```

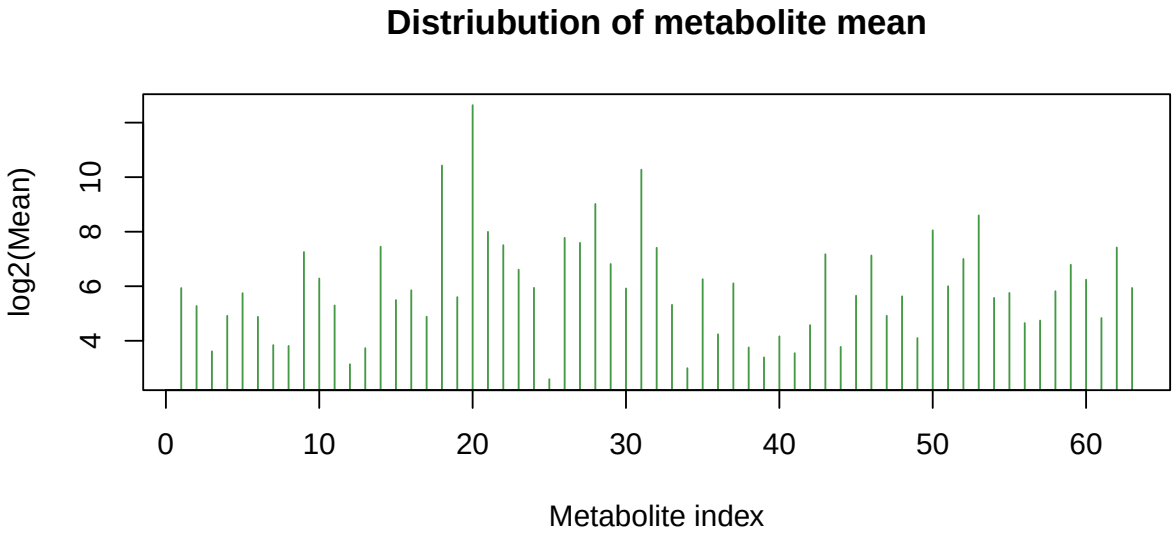


Se puede observar que los metabolitos con los valores medios más altos son la creatinina, el hipurato y el citrato, y que además los niveles de estos 3 metabolitos son más elevados en el caso de los pacientes con caquexia. También puede observarse la presencia de outliers en el boxplot.

Para solucionar todos los problemas mencionados anteriormente se aplica una normalización logarítmica a los datos y se comprueban los resultados.

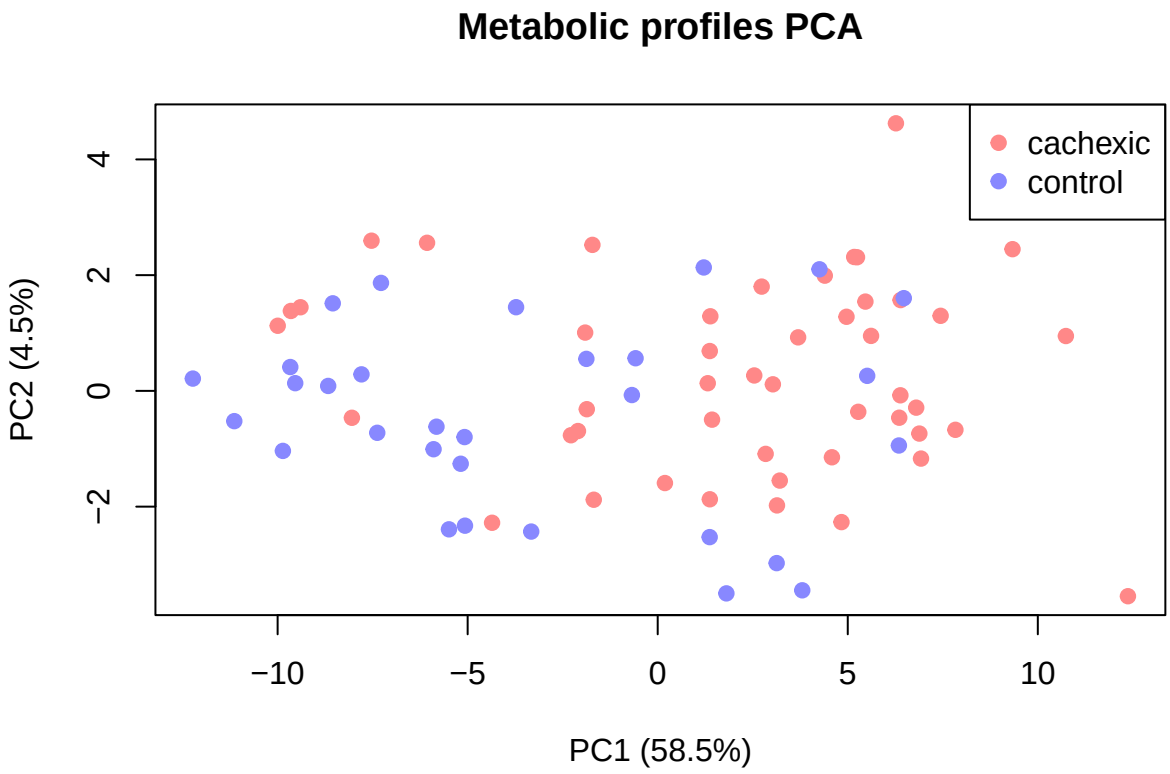
```
# normalización logarítmica
log_matrix <- log2(filtered_matrix + 1)
```





Tras aplicar una transformación logarítmica puede observarse que los valores no están acumulados cerca del 0, si no que se distribuyen de una forma más homogénea. Las magnitudes de los metabolitos se han igualado bastante, esto evitará que unos pocos metabolitos tengan dominancia en el PCA posterior.

Análisis de componentes principales (PCA)



Tras hacer el PCA, se puede observar que hay 2 componentes principales, PC1 y PC2. PC1 explica un 58.5% de la variabilidad total, mientras que la segunda componente explica un 4.5% adicional. La PC1 explica más de un 50% de la variabilidad total, por lo que se deduce que puede existir un patrón que explica más de la mitad del perfil metabólico global. Adicionalmente, se puede observar que hay una separación parcial entre los 2 grupos y que la PC1 capta diferencias metabólicas asociadas al estado caquéxico. A continuación se comprobará que metabolitos contribuyen más a cada una de las componentes.

```
## Metabolitos que más contribuyen a PC1:

##    Creatinine        Valine       Alanine     Glutamine Pyroglutamate
##        0.1564        0.1548        0.1545        0.1529        0.1508
## Dimethylamine cis.Aconitate  Ethanolamine    Asparagine        Serine
##        0.1504        0.1498        0.1497        0.1485        0.1478

##
## Metabolitos que más contribuyen a PC2:

##            Acetate            Sucrose    X2.Oxoglutarate    Methylguanidine
##             0.3593             0.2760            -0.2577            -0.2280
##            Acetone    trans.Aconitate pi.Methylhistidine          Succinate
##            -0.1925             0.1905            -0.1864             0.1854
##            Glucose             Xylose
##             0.1806             0.1760
```

Según los resultados, la creatinina es el metabolito que más contribuye a la PC1, esto es consistente con las observaciones que se hicieron en el análisis exploratorio de los datos. Después de la creatinina, los metabolitos que más pacrecen contribuir a la PC1 son la valina, la alanina y la glutamina. Por otra parte los metabolitos que más contribuyen a la PC2 son el acetato, la sacarosa y el X2.oxoglutarato. La contribución de la creatinina a la PC1 es coherente, ya que este metabolito participa en el metabolismo muscular, por otra parte, el hecho de que sus niveles estén elevados puede indicar que hay procesos de degradación muscular, lo cual es característico de la caquexia.

Discusión

Conclusiones

Referencias

Fearon, K., Strasser, F., Anker, S. D., Bosaeus, I., Bruera, E., Fainsinger, R. L., ... & Baracos, V. E. (2011). Definition and classification of cancer cachexia: an international consensus. *The Lancet Oncology, 12*(5), 489–495. https://doi.org/10.1016/S1470-2045(10)70218-7