Análisis de Datos Ómicos - PEC 1

David López Blanco

29 de marzo, 2025

Contents

Análisis de datos ómicos (M0-157) - Primera prueba de evaluación continua.

Repositorio de GitHub: https://github.com/davidlopezbl/Lopez-Blanco-David-PEC1

Abstract

Este trabajo introduce el análisis de datos ómicos mediante la aplicación de herramientas fundamentales en bioinformática, con un enfoque en `Bioconductor` y la exploración multivariante de datos. Se ha construido un objeto de clase `SummarizedExperiment` a partir de un dataset metabolómico, permitiendo organizar y gestionar los datos de manera estructurada. Además, se han comparado las clases `ExpressionSet` y `SummarizedExperiment`, destacando sus diferencias en términos de flexibilidad, integración de metadatos y compatibilidad con tecnologías modernas. Finalmente, se ha realizado un análisis exploratorio de los datos, aplicando herramientas estadísticas para identificar patrones y tendencias relevantes. Este ejercicio ha permitido consolidar conceptos clave sobre la gestión y exploración de datos ómicos en entornos de análisis bioinformático.

Objetivos

El objetivo principal de este trabajo es aplicar los conocimientos adquiridos sobre el análisis de datos ómicos mediante el uso de herramientas bioinformáticas, con un enfoque en Bioconductor y la exploración multivariante. Para ello, se plantean los siguientes objetivos específicos:

- **Construcción de un objeto `SummarizedExperiment`** a partir de un dataset metabolómico, asegurando una organización adecuada de los datos y su compatibilidad con flujos de trabajo en R.

- **Comparación entre las clases `ExpressionSet` y `SummarizedExperiment`**, destacando sus diferencias en términos de estructura, flexibilidad y aplicación en análisis de datos ómicos.

- **Análisis exploratorio de los datos metabolómicos**, utilizando herramientas estadísticas y métodos de visualización para identificar patrones y tendencias relevantes.

- **Reforzar el uso de herramientas de control de versiones (Git/GitHub)** en el contexto de un análisis bioinformático reproducible.

Métodos

El conjunto de datos utilizado en este análisis, **"human_cachexia.csv"**, proviene de un estudio sobre la caquexia, un síndrome metabólico complejo asociado a enfermedades como el cáncer y caracterizado por pérdida de masa muscular (Evans et al., 2008). Se compone de **77 muestras de orina**, de las cuales **47 corresponden a pacientes con caquexia y 30 a controles sanos**.

El análisis se ha realizado utilizando **R** y el ecosistema **Bioconductor**, que proporciona herramientas especializadas para la manipulación y exploración de datos ómicos. Las principales herramientas y paquetes utilizados incluyen:

- **Bioconductor**: Para la gestión de datos ómicos.

- `SummarizedExperiment` (paquete `SummarizedExperiment`): Para almacenar y estructurar los datos de manera eficiente.

- **Exploración de datos**: Métodos estadísticos y gráficos con `ggplot2` y `heatmap`.
- **Control de versiones**: Uso de **Git** y **GitHub** para asegurar la trazabilidad del análisis.

El análisi ha seguido la siguiente pipeline: 1. **Carga y preprocesamiento de datos**
- Importación del dataset `"human_cachexia.csv"` y verificación de su estructura.
- Conversión de los datos en un objeto **`SummarizedExperiment`**, permitiendo una gestión estructurada de la información.
- Búsqueda de valores faltantes.

2. **Comparación de `ExpressionSet` y `SummarizedExperiment`**
    - Evaluación de las diferencias clave entre ambas clases en cuanto a almacenamiento y manipulación de datos metabolómicos.

    - Análisis de ventajas en términos de flexibilidad y compatibilidad con nuevas tecnologías ómicas.
3. **Análisis exploratorio**
    - **Análisis descriptivo** de las muestras mediante estadísticas básicas.

    - **Visualización de datos** mediante gráficos de distribución, heatmaps y análisis de componentes principales (PCA).

    - **Exploración multivariante** para detectar patrones en la variabilidad de los datos entre grupos de pacientes y controles.

Este enfoque metodológico ha permitido no solo estructurar adecuadamente los datos metabolómicos, sino también aplicar herramientas estadísticas y bioinformáticas para su exploración, proporcionando una visión más clara de las diferencias entre pacientes con caquexia y controles sanos.

Resultados

Ejercicio 1

1. Seleccionad y descargad un dataset de metabolómica, que podéis obtener de metabolomicsWorkbench o del repositorio nutrimetabolomics/metaboData de GitHub.

```
cachexia <- read.csv("human_cachexia.csv", sep=",")
```

Ejercicio 2

2. Cread un objeto de clase `SummarizedExperiment` que contenga los datos y los metadatos (información acerca del dataset, sus filas y columnas). La clase `SummarizedExperiment` es una extensión de ExpressionSet, utilizada por muchas aplicaciones y bases de datos (como es el caso de `metabolomicsWorkbench`). ¿Cuáles son sus principales diferencias con la clase `ExpressionSet`?

```
# Assays (Datos experimentales)
assay_data <- as.matrix(cachexia[, 3:65])
rownames(assay_data) <- cachexia[, 1]

# 'Columns'(sample) data (Metadatos de los pacientes)
col_data <- data.frame(
  cachexic_status = cachexia[, 2],
  row.names = cachexia[, 1]  # Patient IDs
)

# 'Row' (regions-of-interest) data (Metadatos de los features)
row_data <- data.frame(
  metabolite_id = colnames(cachexia)[3:65],
  row.names = colnames(cachexia)[3:65]
)

# Creación del objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(metabolomics = t(assay_data)),
  colData = col_data,
  rowData = row_data
)

# Mostrar SummarizedExperiment creado
se
```

```
class: SummarizedExperiment
dim: 63 77
metadata(0):
assays(1): metabolomics
rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
  pi.Methylhistidine tau.Methylhistidine
rowData names(1): metabolite_id
colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
colData names(1): cachexic_status
```

Para poder gestionar estrructuras de datos generadas por las nuevas tecnologías de alto rendimiento en biología molecular ha sido necesario el desarrollo paralelo de herramientas bioinformáticas apropiadas, como el tradicional `ExpressionSet` y el moderno `SummarizedExperiment` (de Bioconductor).
La clase `ExpressionSet`[STHDA, 2024], presente en la libería `Biobase`, se creó durante la era de los microarrays para poder cargar datos de expresión génica, típicamente representados como una matriz de intensidad única. Esta clase funciona como una estructura matricial aumentada con tres componentes principales: la matriz de expresión (`exprs()`), datos fenotípicos (`pData()`) y anotaciones de características (`fData()`). Este diseño permitió y permite utilizar pipelines de trabajo genómico y transcriptómico de manera eficaz donde los resultados experimentales eran inherentemente bidimensionales y las coordenadas genómicas eran innecesarias.
Por el contrario, la clase `SummarizedExperiment` [Team, 2024] se ha desarrolado para abordar las complejidades introducidas por las tecnologías de secuenciación de nueva generación. La clase va más allá de un simple paradigma matriz-contenedor, con el uso cuatro innovaciones clave: compatibilidad con múltiples matrices de ensayo, integración nativa de coordenadas genómicas mediante objetos GRanges, contenedores de metadatos flexibles basados en DataFrame y arquitectura de clase S4 optimizada para grandes conjuntos de datos.
El hecho de que actualmente sea más común hacer secuenciación en vez de usar microarrays ha dejado ver limitaciones en el uso de `ExpressionSet` como:

1. Compatibilidad con múltiples matrices de ensayo, lo que permite almacenar recuentos brutos, valores normalizados y datos transformados como ensayos separados dentro de un mismo objeto.
2. Integración nativa de coordenadas genómicas a través del uso de objetos `GRanges`, permitiendo un análisis más completo de los datos, especialmente en estudios que relacionan metabolitos con sus reguladores genéticos.
3. Uso de contenedores de metadatos flexibles basados en `DataFrame`, lo que facilita la gestión de datos heterogéneos, como variables clínicas, covariables técnicas y anotaciones de características.
4. Arquitectura de clase S4, que permite la optimización y escalabilidad para trabajar con grandes volúmenes de datos.

Si bien `ExpressionSet` sigue siendo adecuado para análisis sencillos de tipo microarray, `SummarizedExperiment` se ha convertido en el estándar de facto para la investigación metabolómica moderna, entre otras ómicas. Su diseño anticipa la creciente complejidad de los datos ómicos, manteniendo la retrocompatibilidad. Esta transición a `SummarizedExperiment` refleja la evolución más amplia de la bioinformática, desde la integración monoómica a la multimodal, lo que permite aprovechar al máximo las tecnologías emergentes en metabolómica espacial, análisis de single-cell y/o estudios de biología de sistemas.

Ejercicio 3

3. Llevad a cabo un análisis exploratorio que os proporcione una visión general del dataset en la línea de lo que hemos visto en las actividades de este reto.
La primera parte de este análisis exploratorio consite en entender la estructura del dataset de estudio, incluyendo la dimensión de este, así como las variables recogidas. También es interesante conocer si falta algún valor (missing). Otra forma de tener una idea general/exploratoria de los datos es mostrar la cabecera de los datos (`head(cachexia)`), pero no se realiza en este estudio porque simplemente resultaría en un volcado de los datos.

```
# Dimensión dataset
dim(cachexia)
```

```
[1] 77 65
```

```r
# Variables del dataset
colnames(cachexia) %>%
  paste(collapse = ", ") %>%
  cat()
```

```
Patient.ID, Muscle.loss, X1.6.Anhydro.beta.D.glucose, X1.Methylnicotinamide, X2.Aminobutyrate, X2.Hydro
```

```r
# Contaje pacientes de cada estado de "Muscle.loss"
table(cachexia[[2]])
```

```
cachexic  control
      47       30
```

```r
# Cálculo missing values
contaje_missing <- sum(colSums(is.na(cachexia[, 3:65])) > 0)
cat("Metabolitos con missing values:", contaje_missing, "/", ncol(cachexia[, 3:65]))
```

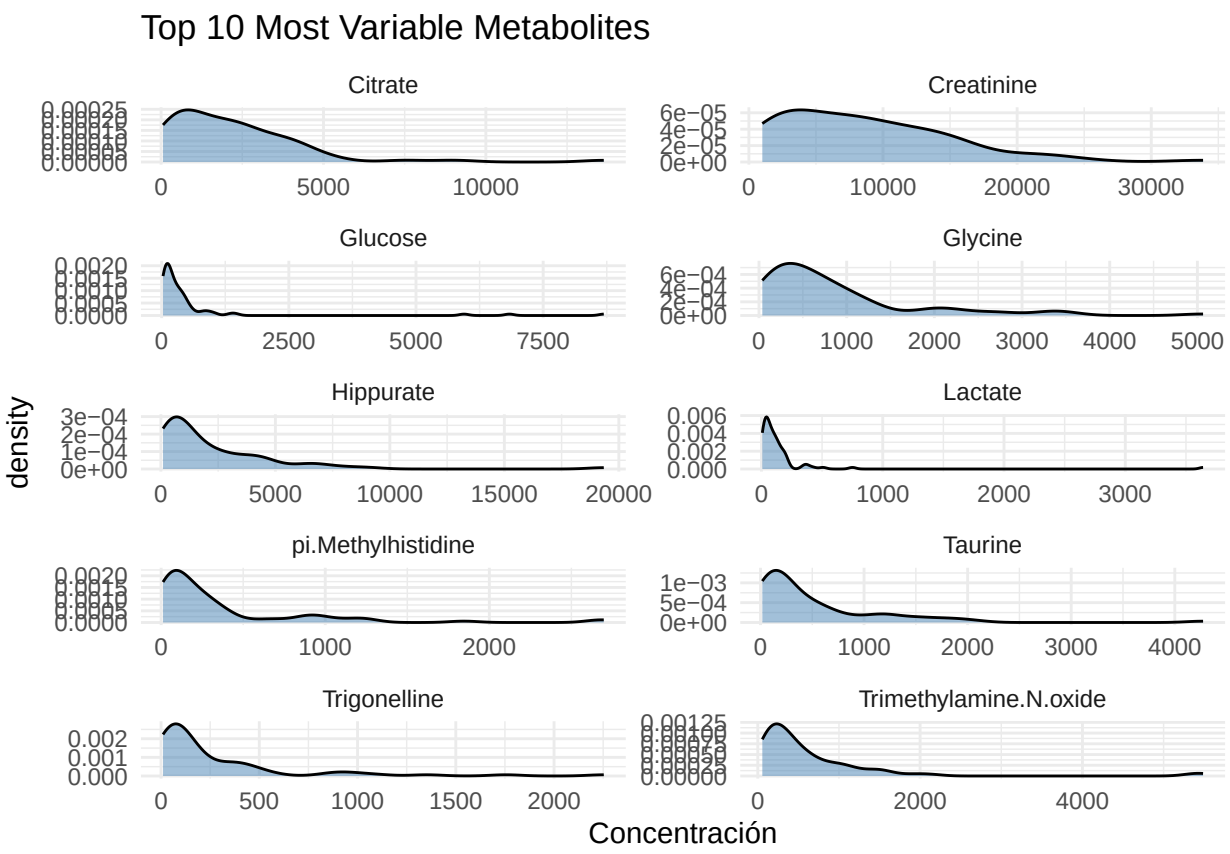
```
Metabolitos con missing values: 0 / 63
```

Una vez se conoce la estructura general de los datos, se osberva como en la primera columna está el ID de los 77 pacientes y en la segunda está el factor `Muscle.loss`, el cual nos informa del estado del paciente (caquexia o no). Tal y como indica la descripción del dataset, este dataset incluye 47 pacientes caquéxicos y 30 pacientes control.
Por otro lado, de la columna 3 a la 65 hay recogidos valores de concentración de 63 metabolitos en orina.
Para poder entender mejor los datos, se propone visualizar los valores del dataset con gráficos de densidad y boxplot de los valores.

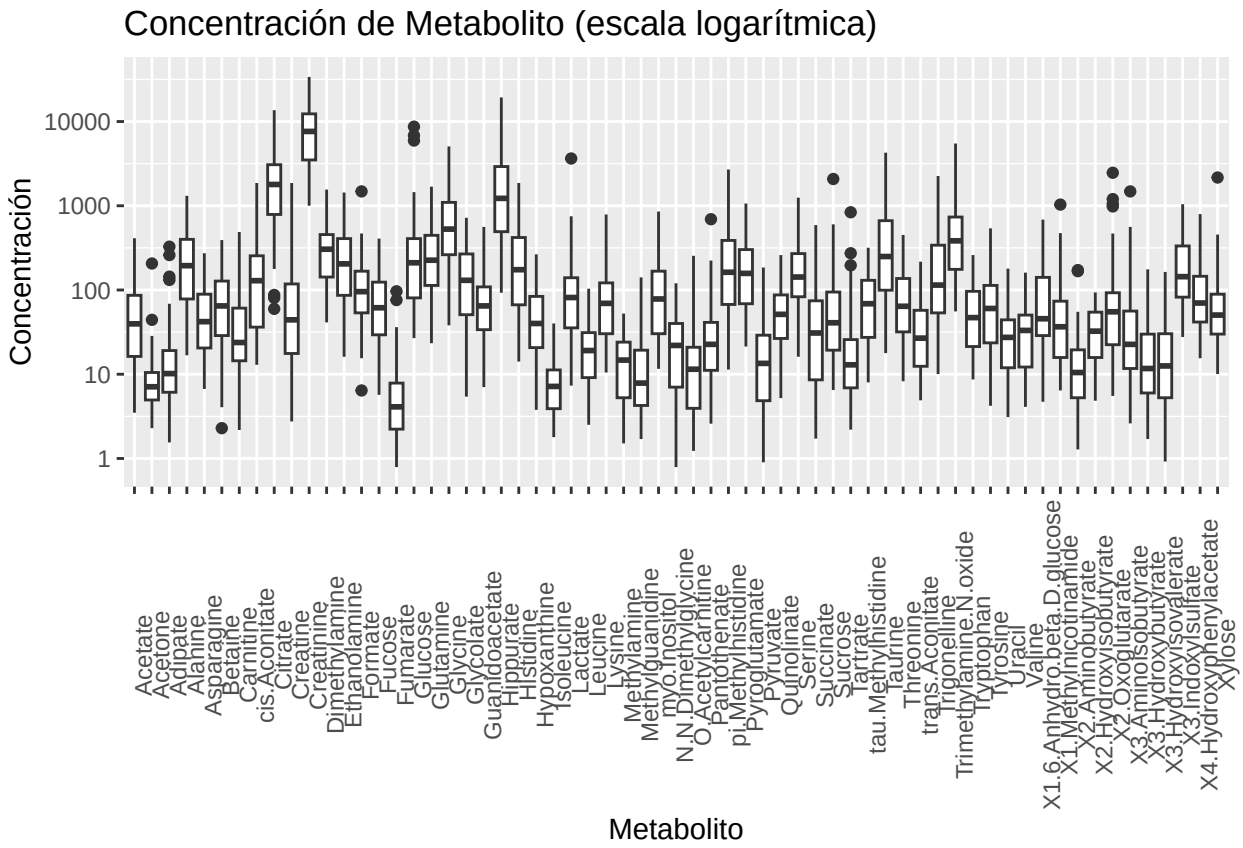
```r
# Reesstructuración de los datos de su visualización
cachexia_melted <- cachexia |>
  pivot_longer(
    cols = 3:65,
    names_to = "Metabolito",
    values_to = "Concentración"
  ) |>
  select(Patient.ID, Muscle.loss, Metabolito, Concentración)

# Top 10 metabolitos más variables
top_metabolitos <- cachexia_melted %>%
  group_by(Metabolito) %>%
  summarise(variance = var(Concentración, na.rm = TRUE)) %>%
  slice_max(variance, n = 10) %>%
  pull(Metabolito)

# Gráfico densidad de este top de metabolitos
cachexia_melted %>%
  filter(Metabolito %in% top_metabolitos) %>%
  ggplot(aes(Concentración)) +
  geom_density(fill = "steelblue", alpha = 0.5) +
  facet_wrap(~Metabolito, scales = "free", ncol = 2) +  # 2-column layout
  theme_minimal() +
  labs(title = "Top 10 Most Variable Metabolites")
```



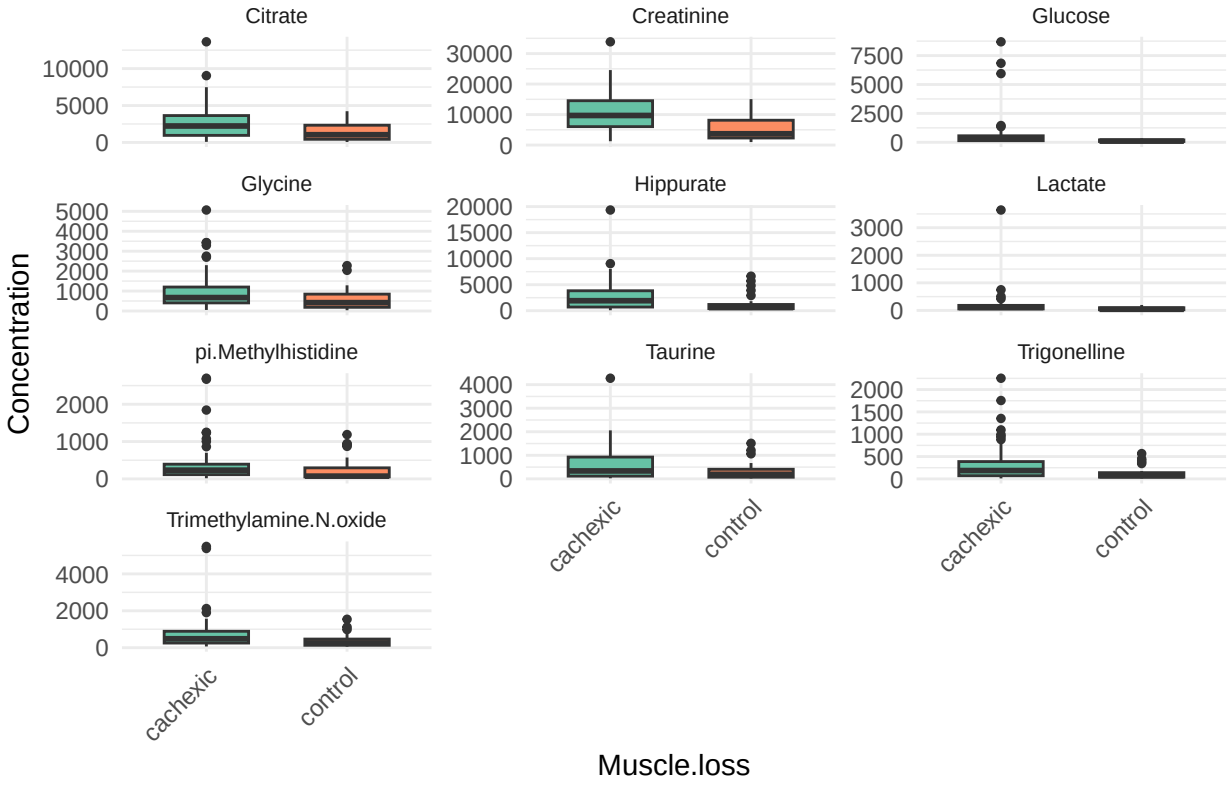

```r
# Boxplots (escala logarítmica)
ggplot(cachexia_melted, aes(x = Metabolito, y = Concentración)) +
  geom_boxplot() +
  scale_y_log10() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Concentración de Metabolito (escala logarítmica)")
```



Otro estudio interesante que se puede hacer de estos datos es conocer y calcular la correlación entre los diferentes metabolitos. Esto nos ayuda a conocer mejor si hay posibles dependencias entre las diferentes medidas. Aunque no se muestra el resultado en este informe, esto se puede realizar usando la función `cor` con *pairwise.complete.obs* y después representarlo usando `heatmap`.

Finalmente, para comprender mejor las diferencias entre grupos se puede hacer un boxplot comparativo para cada aminoácido. También puede ser interesante hacer un PCA de los datos para entender mejor la clasificaciones a partir de los datos experimentales.

```
# Comparación concentración de metabolitos entre niveles
ggplot(cachexia_melted %>%
         filter(Metabolito %in% top_metabolitos),
       aes(x = Muscle.loss, y = Concentración, fill = Muscle.loss)) +
  geom_boxplot(width = 0.6, outlier.size = 1) +
  facet_wrap(~Metabolito, scales = "free_y", ncol = 3) +
  scale_fill_brewer(palette = "Set2") +
  labs(title = "Comparativa Concentración Metabolitos en función Estado de Caquexia",
       y = "Concentration") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1),
        strip.text = element_text(size = 8))
```



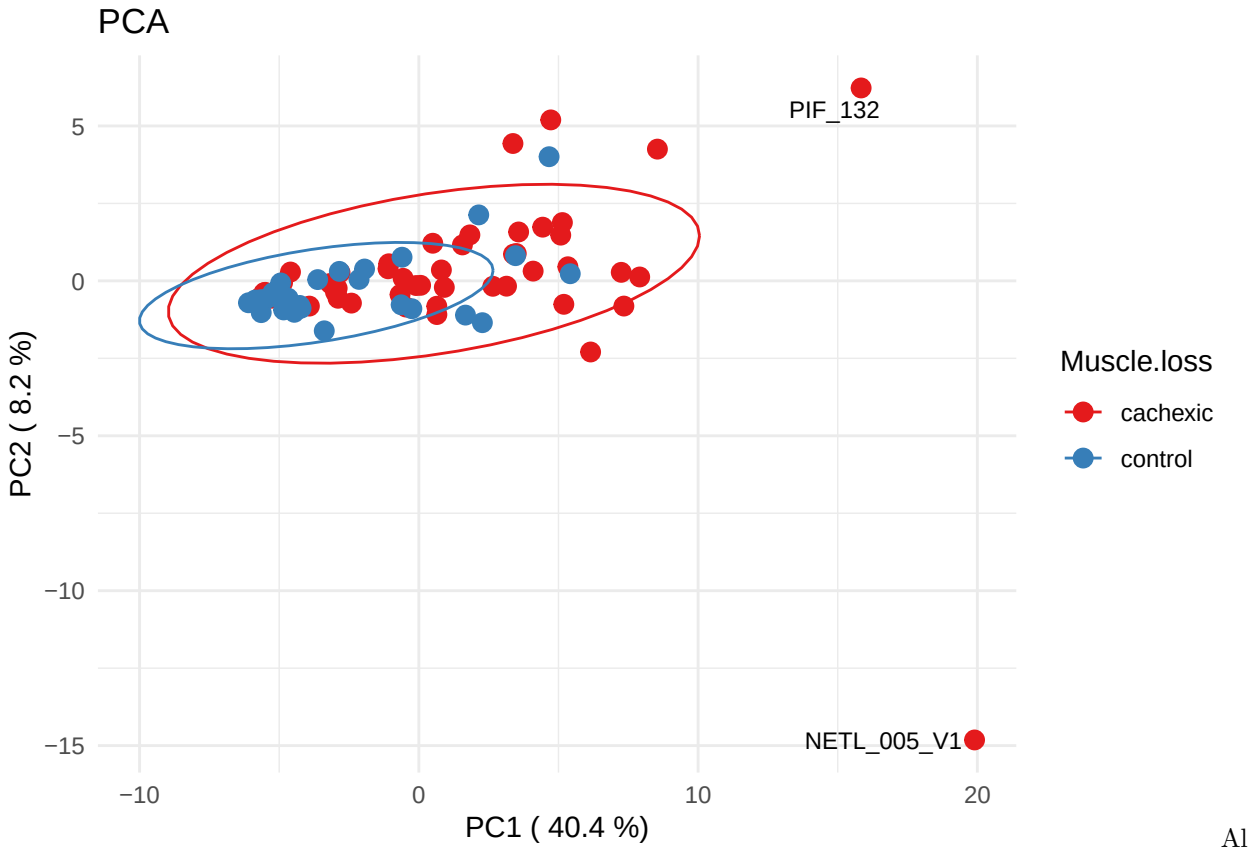

```r
# Estudio PCA de los datos
pca_resultados <- prcomp(cachexia[, 3:65], scale. = TRUE)

# Creación de un dataframe con los datos del PCA
pca_data <- data.frame(
  PC1 = pca_resultados$x[,1],
  PC2 = pca_resultados$x[,2],
  Muscle.loss = as.factor(cachexia$Muscle.loss),
  Patient.ID = cachexia$Patient.ID
)

# Cálculo de la variancia
var_explained <- round(100 * pca_resultados$sdev^2 / sum(pca_resultados$sdev^2), 1)

# Create the plot
ggplot(pca_data, aes(PC1, PC2, color = Muscle.loss)) +
  geom_point(size = 3) +
  ggrepel::geom_text_repel(
    data = subset(pca_data, abs(scale(PC1)) > 2.5),
    aes(label = Patient.ID), color = "black", size = 3
  ) +
  stat_ellipse() +
  labs(
    x = paste("PC1 (", var_explained[1], "%)"),
    y = paste("PC2 (", var_explained[2], "%)"),
    title = "PCA"
  ) +
```

```
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```



Al hacer el PCA de los datos se observa como no hay una muy buena separación de estos. Esto podría ser al hecho de que se osberva que hay un par de valores que podrían considerarse outliers que sería interesante eliminar (si realmetne lo son), y entonces volver a representar. No obstante, como se está realizando un estudio exploratorio esto no se realizará.

Discusión

El análisis realizado ha permitido aplicar herramientas bioinformáticas para la exploración de datos metabolómicos en el contexto de la caquexia. A través de la construcción de un objeto `SummarizedExperiment` y el análisis exploratorio de los datos, se han abordado algunas de las metodologías y herramientas bioinformáticas que resultan esenciales para entender las complejidades de este síndrome. No obstante, cabe destacar algunas de las limitaciones del trabajo y su relevanica para el estudio de la enfermedad estudiada con este dataset, la caquexia.

Entre las limitaciones se encuentran:

1. **Falta de variables clínicas adicionales**: Aunque el dataset distingue entre pacientes con caquexia y controles, no se dispone de información adicional relevante como características clínicas detalladas (edad, sexo, comorbilidades), tratamientos previos, o variables de estilo de vida (por ejemplo, dieta o actividad física). Estas variables podrían influir significativamente en los perfiles metabolómicos y deberían ser consideradas en futuras investigaciones para obtener una comprensión más completa de la caquexia.

2. **Tamaño muestral limitado**: Aunque el estudio inlcuye un total de 77 muestras, podría llegar a considerarse que esu un número relativamente pequeño, cosa que podría afectar a las conclusiones del estudio y al posible desarrollo de heramientas calsificatorias a partir de los datos.

3. **Enfoque exploratorio limitado**: Este trabajo se centra principalmente en la exploración descriptiva y la visualización multivariante de los datos metabolómicos. Aunque estas herramientas son fundamentales para detectar patrones generales, no se ha abordado un análisis más profundo, como el modelado predictivo o la identificación de biomarcadores específicos que puedan servir como indicadores de la progresión de la caquexia.Cabe decir que tampoco se ha hecho estudios de outliers propiamente para mejorar la fiabilidad de los datos y sus posibles extrapolaciones.

4. **Preprocesamiento y normalización de datos**: El hecho de tener los datos de concentración directamente sin información de los métodos/técnicas de normalización y preprocesamiento limita el entendiemiento del orgien de los datos. Idealmente, esto se debería conocer y, de hecho, se podría estudiar diferentes técnicas para poder analizar más correctamente los datos brutos.

Respecto a la relevancia de este estudio, se conoce que la caquexia es un síndrome metabólico complejo que afecta a una gran proporción de pacientes con enfermedades crónicas, como el cáncer, y se caracteriza por una pérdida de masa muscular y, en algunos casos, pérdida de grasa. Por otro lado, la metabolómica estudia el metabolismo global presentes en el organismo, lo que ofrece una herramienta poderosa para identificar biomarcadores asociados con la caquexia. Por eso resulta interesante el uso de esta ómica para el análisis. No obstante, a pesar de las ventajas de estas herramientas, es importante considerar que la metabolómica es solo una parte de la historia. Para obtener una comprensión más completa de la caquexia, sería necesario integrar datos de otras capas ómicas, como la genómica y la transcriptómica, que podrían aportar información sobre los mecanismos moleculares subyacentes a la enfermedad.

Este análisis ha demostrado que el uso de herramientas como `SummarizedExperiment` es valioso para gestionar datos metabolómicos de manera eficiente. La capacidad de esta clase para manejar múltiples matrices de datos y metadatos asociados facilita la integración de diferentes tipos de información y la realización de análisis multivariantes, fundamentales para entender la complejidad biológica de la caquexia.

Entre algunas de las posibles mejoras o futuras direcciones para estudios posteriroes se propone: ampliar el tamaño muestra (una mayor n siempre mejora y hace más robusto el análisis, desarrollo de herramientas y la obtención de patrones), integrar los datos ómicos con el historial clínico de los pacientes, y utilizar modelos predictivos y *machine leatning* para poder desarrollar herramientas de detección temprana.

Conclusiones

En resumen, este estudio ha proporcionado una introducción valiosa a las herramientas bioinformáticas y estadísticas para el análisis de datos metabolómicos en caquexia. Si bien las limitaciones mencionadas deben ser consideradas, el trabajo realizado sienta las bases para investigaciones más profundas y sofisticadas que podrían contribuir al desarrollo de nuevos biomarcadores y tratamientos para la caquexia. A medida que los métodos de análisis y las tecnologías ómicas continúan avanzando, se espera que los estudios en esta área puedan proporcionar información crucial para el manejo de esta condición debilitante.

Referencias

STHDA. *ExpressionSet and SummarizedExperiment*, 2024. URL https://www.sthda.com/english/wiki/expressionset-and-summarizedexperiment. Accessed: 2024-06-20.

Bioconductor Core Team. *SummarizedExperiment: A Bioconductor Container for Summarized Experiments*, 2024. URL https://www.bioconductor.org/packages/devel/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html. Bioconductor package version 3.19.