Ivanov_Varsamov_Nikolay_pec1

Nikolay Ivanov

2025-04-02

Índice

Introducción

Selección del dataset

Procesamiento de los datos

Filtrado y preparación del objeto `SummarizedExperiment`

Análisis exploratorio

Análisis de componentes principales (PCA)

Heatmap de features más variables

Clustering jerárquico de muestras

Interpretación de resultados

Referencias

Introducción

Se ha llevado a cabo un análisis exploratorio de datos metabolómicos obtenidos de un proyecto de metabolómica (mediante espectrometría de masas en tandem con cromatografía líquida LC-MS) durante el proceso de malteado de cebada. A partir de los datos crudos y los metadatos proporcionados, se construyó un objeto SummarizedExperiment utilizando herramientas del ecosistema Bioconductor, permitiendo integrar las intensidades metabolómicas con la información experimental. El análisis incluyó una transformación logarítmica, normalización y selección de las features más variables. Posteriormente, se aplicaron métodos de reducción de dimensionalidad (PCA), visualización mediante mapas de calor y clustering jerárquico.

Repositorio GIT de depósito:

https://github.com/nmi291434/Ivanov_Varsamov_Nikolay_PEC1

Selección del dataset

Se ha seleccionado el dataset ST003289, de la página metabolomicsworkbench

Procesamiento de los datos

Observamos la hoja de metadatos disponible

```
library(SummarizedExperiment)
```

```
## Cargando paquete requerido: MatrixGenerics
```

```
## Warning: package 'MatrixGenerics' was built under R version 4.4.2
```

```
## Cargando paquete requerido: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 4.4.3
```

```
##
## Adjuntando el paquete: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Cargando paquete requerido: GenomicRanges
```

```
## Cargando paquete requerido: stats4
```

```
## Cargando paquete requerido: BiocGenerics
```

```
##
## Adjuntando el paquete: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Cargando paquete requerido: S4Vectors

##
## Adjuntando el paquete: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Cargando paquete requerido: IRanges

## Warning: package 'IRanges' was built under R version 4.4.2

##
## Adjuntando el paquete: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Cargando paquete requerido: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.4.2

## Cargando paquete requerido: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Adjuntando el paquete: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
library(tibble)
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following object is masked from 'package:Biobase':
##
##     combine

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.3
```

```r
# Ver hojas disponibles
excel_sheets("Study design.xlsx")
```

```
## [1] "Hoja1"
```

```r
# Leer la hoja única del Excel
metadata <- read_excel("Study design.xlsx")

# Filtrar solo las muestras de plataforma LC-MS
metadata_lcms <- metadata %>%
  filter(Platform == "LC-MS")

# Ver la matriz resultante
print(metadata_lcms)
```

```
## # A tibble: 67 x 6
##    mb_sample_id local_sample_id      `Sample source` Genotype Treatment Platform
##    <chr>        <chr>                <chr>           <chr>    <chr>     <chr>
##  1 SA356372     22-1-blank-blank-Bl~ Seed            Conrad   BLANK     LC-MS
##  2 SA356373     39-1-blank-blank-Bl~ Seed            Conrad   BLANK     LC-MS
##  3 SA356374     55-1-blank-blank-Bl~ Seed            Conrad   BLANK     LC-MS
##  4 SA356384     63-1-2-dog0-11       Seed            Conrad   DOG0      LC-MS
##  5 SA356385     16-1-2-dog0-10       Seed            Conrad   DOG0      LC-MS
##  6 SA356386     18-1-2-dog0-17       Seed            Conrad   DOG0      LC-MS
##  7 SA356387     48-1-2-dog0-16       Seed            Conrad   DOG0      LC-MS
##  8 SA356388     64-1-2-dog0-13       Seed            Conrad   DOG0      LC-MS
##  9 SA356389     13-1-2-dog0-14       Seed            Conrad   DOG0      LC-MS
## 10 SA356390     27-1-2-dog0-15       Seed            Conrad   DOG0      LC-MS
## # i 57 more rows
```

```r
# Si quieres convertirlo a una matriz R base (opcional)
metadata_matrix <- as.matrix(metadata_lcms)
```

Confirmamos la unicalidad de los local_sample_id empleados en el proyecto

```r
anyDuplicated(metadata_lcms$local_sample_id)
```

```
## [1] 0
```

Abrimos archivo de resultados de lcms

```r
lcms_data <- read.delim("ST003289_AN005386_Results (1).txt", check.names = FALSE)
str(lcms_data)
```

```
## 'data.frame':    8921 obs. of  55 variables:
##  $ Primary adduct_Rt: chr  "105.071_543.9" "184.394_500.6" "186.874_335.2" "184.142_437.2" ...
##  $ 2-1-4-dog3-29    : num  2.44 2.78 2.83 2.92 3.06 ...
##  $ 3-1-6-kilned-46  : num  20.05 2.97 2.8 3.36 2.85 ...
##  $ 4-1-3-dog1-20    : num  8.58 10.88 4.13 4.43 3.36 ...
##  $ 5-1-4-dog3-32    : num  9.46 3.05 2.39 3.53 2.29 ...
```

```
##  $ 6-1-6-kilned-51  : num  13.93 4.22 6.71 3.4 3.33 ...
##  $ 7-1-1-dry-5      : num  14.8 2.7 3.23 3.53 1.68 ...
##  $ 9-1-6-kilned-52  : num  5.6 3.66 6.61 3.5 6.16 ...
##  $ 10-1-4-dog3-36   : num  3.02 7.74 3.76 3.31 7.28 ...
##  $ 11-1-5-dog5-42   : num  5.11 6.38 5.26 3.66 4.3 ...
##  $ 12-1-6-kilned-54 : num  27.25 3.16 2.52 2.7 3.47 ...
##  $ 13-1-2-dog0-14   : num  7.16 3.09 6.48 2.78 5.54 ...
##  $ 14-1-3-dog1-22   : num  13.05 3.03 2.22 4.1 3.45 ...
##  $ 16-1-2-dog0-10   : num  14.86 3.58 3.08 10.8 3.65 ...
##  $ 17-1-4-dog3-30   : num  30.42 3.75 2.53 6.69 3.84 ...
##  $ 18-1-2-dog0-17   : num  16.08 4.41 7.21 3.13 3.17 ...
##  $ 19-1-5-dog5-45   : num  24.91 3.01 5.6 2.97 3.04 ...
##  $ 20-1-3-dog1-26   : num  16.97 3.43 1.97 4.78 3.15 ...
##  $ 21-1-3-dog1-23   : num  8.29 3.93 11.64 5.6 6.25 ...
##  $ 24-1-5-dog5-37   : num  5.79 4.35 7.53 4.14 2.78 ...
##  $ 25-1-4-dog3-31   : num  23.13 4.5 11.79 2.61 4.68 ...
##  $ 26-1-5-dog5-38   : num  10.39 3.46 3.12 5.15 3.95 ...
##  $ 27-1-2-dog0-15   : num  32.98 4.76 2.86 3.56 4.12 ...
##  $ 28-1-3-dog1-21   : num  2.02 4.86 3.7 3.94 7.34 ...
##  $ 29-1-3-dog1-24   : num  6.44 3.11 8.33 3.79 3.43 ...
##  $ 31-1-6-kilned-53 : num  39.84 4.23 37.11 3.4 2.71 ...
##  $ 32-1-6-kilned-50 : num  21.84 2.05 5.7 4.43 2.72 ...
##  $ 33-1-6-kilned-47 : num  2.91 3.14 3.36 3.81 2.42 ...
##  $ 34-1-6-kilned-48 : num  3.95 3.19 11.72 3.23 3.7 ...
##  $ 35-1-1-dry-4     : num  18.81 2.18 2.5 2.26 3.78 ...
##  $ 36-1-4-dog3-33   : num  14.89 3.41 11.41 2.64 2.38 ...
##  $ 38-1-5-dog5-40   : num  16.26 4.08 3.27 3.21 3.24 ...
##  $ 40-1-5-dog5-44   : num  4.83 2.86 7.21 3.95 3.15 ...
##  $ 41-1-2-dog0-18   : num  29.81 3.65 8.13 3.04 4.3 ...
##  $ 42-1-4-dog3-28   : num  17.28 4.47 3.89 3.67 4.95 ...
##  $ 43-1-1-dry-2     : num  6.67 3.04 2.76 3.46 3.09 ...
##  $ 44-1-1-dry-8     : num  28.43 3.91 2.69 3.77 1.86 ...
##  $ 46-1-1-dry-9     : num  15.04 2.54 7.08 3.09 3.66 ...
##  $ 47-1-5-dog5-43   : num  17.7 3.48 2.79 3.03 2.42 ...
##  $ 48-1-2-dog0-16   : num  25.9 3.94 3.98 3.62 3.44 ...
##  $ 49-1-1-dry-1     : num  8.31 3.95 6.89 3.4 3.51 ...
##  $ 50-1-1-dry-7     : num  184.21 2.87 6.97 2.97 2.98 ...
##  $ 51-1-3-dog1-27   : num  90.75 2.41 11.81 3.51 3.24 ...
##  $ 53-1-3-dog1-25   : num  10.74 4.17 2.98 3.21 4.36 ...
##  $ 54-1-1-dry-3     : num  1.1 3.22 6.35 3 4.33 ...
##  $ 56-1-2-dog0-12   : num  1.04 2.67 4.18 3.06 3.5 ...
##  $ 57-1-4-dog3-34   : num  1.91 3.51 7.9 2.69 4.37 ...
##  $ 58-1-1-dry-6     : num  9.71 3.18 9.1 3.72 3.51 ...
##  $ 59-1-6-kilned-49 : num  2.21 3.42 6.07 6.35 3.52 ...
##  $ 61-1-3-dog1-19   : num  9.39 3.04 2.66 2.91 3.66 ...
##  $ 62-1-5-dog5-41   : num  19.53 2.67 4.06 4.02 7.99 ...
##  $ 63-1-2-dog0-11   : num  4.55 4.04 3.54 3.79 4.55 ...
##  $ 64-1-2-dog0-13   : num  1.7 3.03 4.11 2.73 4.33 ...
##  $ 65-1-4-dog3-35   : num  16.01 2.89 6.17 2.84 4.02 ...
##  $ 66-1-5-dog5-39   : num  1.02 3.5 2.42 4.93 11.14 ...
```

Verificamos coincidencias

```r
# Extraer nombres de muestra del archivo LC-MS (sin contar la primera columna que es el ID de feature)
sample_names_lcms <- colnames(lcms_data)[-1]

# Verificar cuáles están presentes en los metadatos
matching_samples <- sample_names_lcms %in% metadata_lcms$local_sample_id

# Mostrar resumen
table(matching_samples)
```

```
## matching_samples
## TRUE
##   54
```

Filtrado y preparación del objeto `SummarizedExperiment`

Procedemos a la generación del objeto summarized experiment para los resultados lcms

```r
# Paso 1: Identificar muestras comunes
common_samples <- intersect(metadata_lcms$local_sample_id, colnames(lcms_data)[-1])

# Paso 2: Filtrar datos de intensidades
assay_matrix <- lcms_data[, c("Primary adduct_Rt", common_samples)]
rownames(assay_matrix) <- assay_matrix$`Primary adduct_Rt`
assay_matrix <- assay_matrix[, -1]  # eliminar columna de nombres de features

# Convertir a matriz numérica
assay_matrix <- as.matrix(assay_matrix)

# Paso 3: Filtrar metadatos
col_data <- metadata_lcms %>%
  filter(local_sample_id %in% common_samples) %>%
  arrange(match(local_sample_id, common_samples))  # para mantener el orden

# Paso 4: Crear objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = assay_matrix),
  colData = col_data
)

# Paso 5: Metadatos experimentales
metadata(se)$experiment_info <- list(
  treatment_id = "TR003418",
  treatment_summary = "9 samples were collected (three per micro-malting replicate) at 6 different stag
  chromatography = list(
    chromatography_id = "CH004083",
    instrument_name = "Waters Acquity",
    column = "Waters Acquity UPLC CSH Phenyl Hexyl (1.7 µM, 1.0 × 100 mm)",
    temperature = "65°C",
    flow_rate = "200 µL/min",
    solvents = list(
      A = "100% water; 0.1% ammonium formate",
      B = "100% acetonitrile; 0.1% formic acid"
```

```
    ),
    type = "Reversed phase"
  ),
  ms_analysis = list(
    analysis_id = "AN005386",
    ms_type = "ESI",
    ms_instrument = "Waters Xevo-G2-XS",
    ion_mode = "POSITIVE",
    units = "Peak Area"
  )
)


# Ver resumen
se
```

```
## class: SummarizedExperiment
## dim: 8921 54
## metadata(1): experiment_info
## assays(1): counts
## rownames(8921): 105.071_543.9 184.394_500.6 ... 758.567_473.8
##   782.568_466.5
## rowData names(0):
## colnames(54): 63-1-2-dog0-11 16-1-2-dog0-10 ... 32-1-6-kilned-50
##   33-1-6-kilned-47
## colData names(6): mb_sample_id local_sample_id ... Treatment Platform
```

Guardamos el objeto summarized experiment en formato rda

```
save(se, file = "lcms_summarized.Rda")
```

Análisis exploratorio

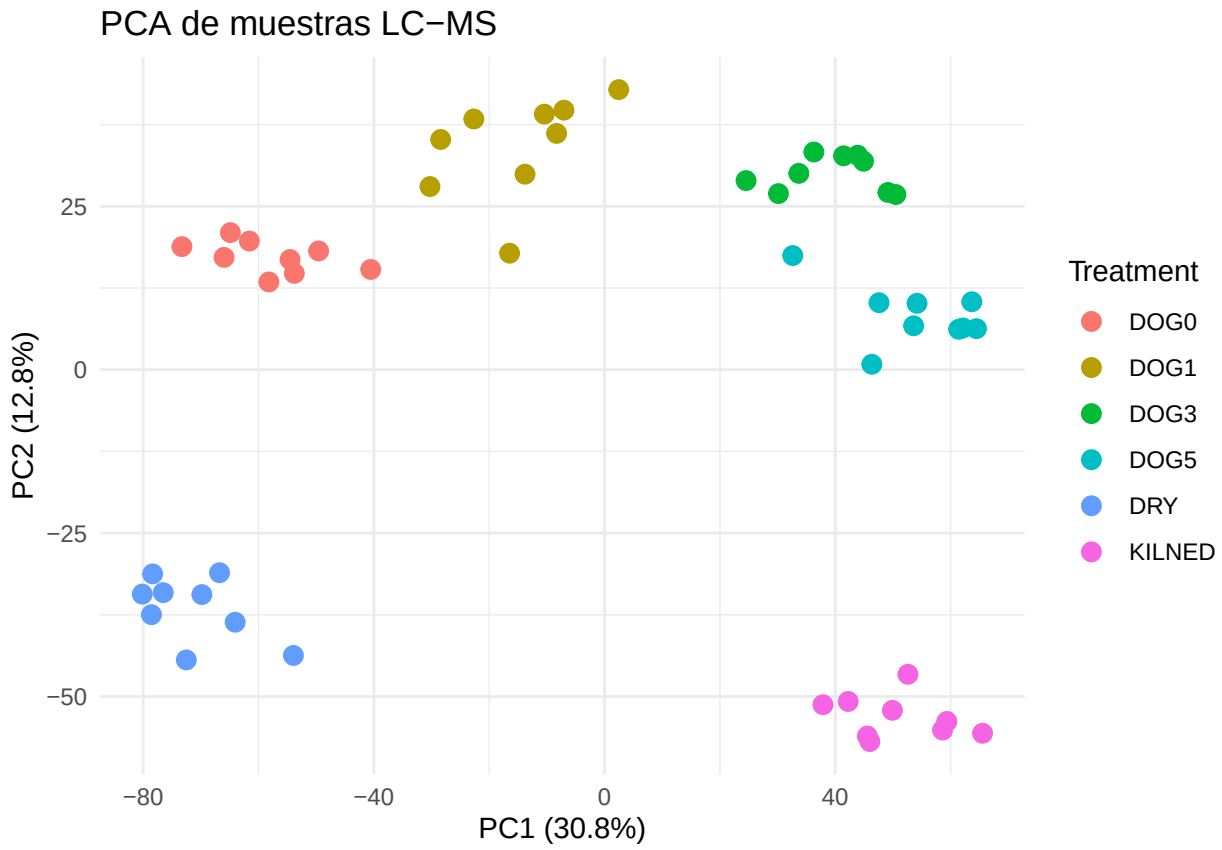
Análisis de componentes principales (PCA)

Realizamos un PCA sobre los datos normalizados

```
# Extraer matriz de intensidades
expr <- assay(se)

# Log-transformación (añadimos un pseudoconteo para evitar log(0))
log_expr <- log2(expr + 1)

# Normalizar (centrar y escalar)
log_expr_scaled <- t(scale(t(log_expr)))

# PCA
pca <- prcomp(t(log_expr_scaled))

# Gráfico
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```r
pca_df <- as.data.frame(pca$x)
pca_df$Treatment <- colData(se)$Treatment

ggplot(pca_df, aes(PC1, PC2, color = Treatment)) +
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "PCA de muestras LC-MS",
       x = paste0("PC1 (", round(summary(pca)$importance[2,1] * 100, 1), "%)"),
       y = paste0("PC2 (", round(summary(pca)$importance[2,2] * 100, 1), "%)"))
```



El PCA obtenido muestra una separación clara entre tratamientos, especialmente KILNED y DRY a lo largo de PC1, que explica el 30,8% de la variabilidad. Las muestras DOG0, DOG1, DOG3 y DOG5 se agrupan más cerca entre sí, con cierta diferenciación en PC2 (12,8%). La distribución sugiere un efecto notable del tratamiento sobre el perfil metabolómico y una buena consistencia entre réplicas.

Heatmap de features más variables

Así pues, seguimos con un heatmap para observar la abundancia de metabolitos según tratamiento y muestra:

```r
# Calcular varianza por fila
feature_vars <- apply(log_expr, 1, var)

# Seleccionar top 50 features más variables
```

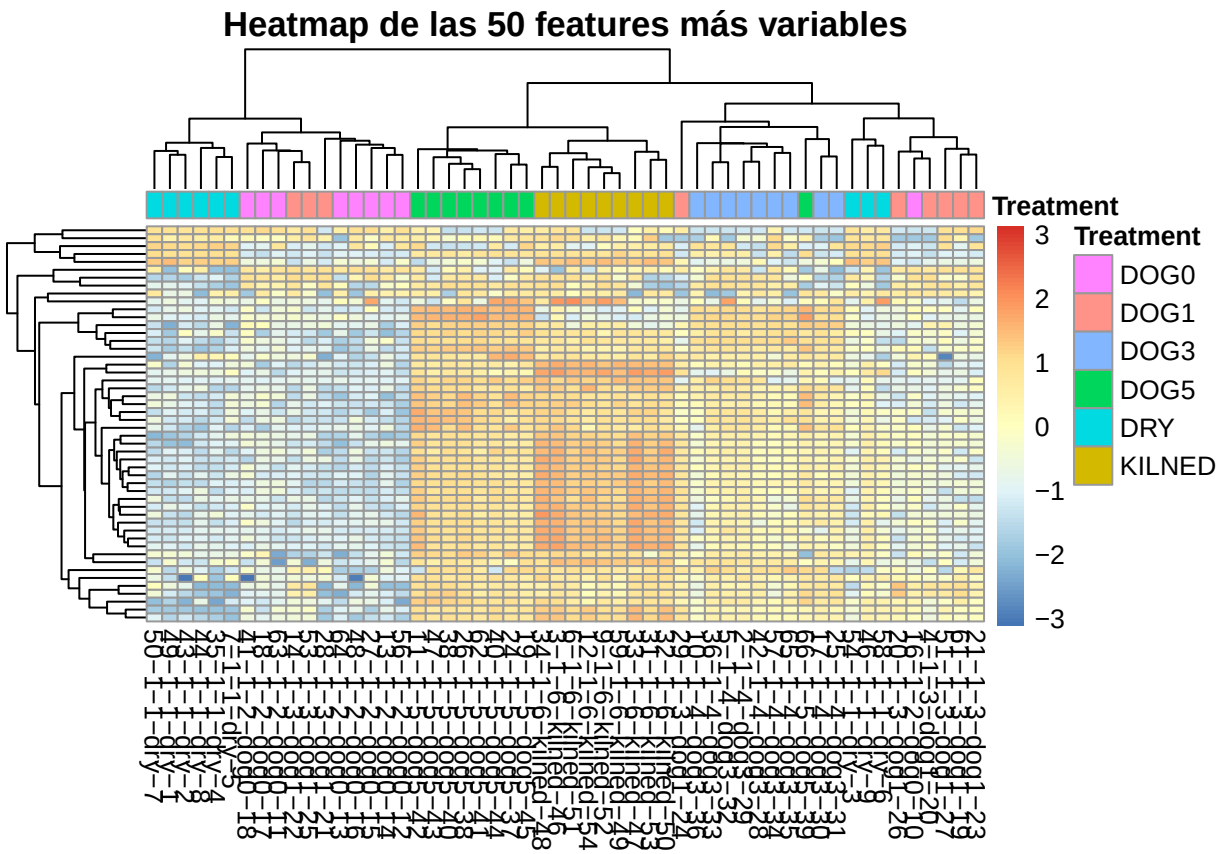
```r
top_features <- names(sort(feature_vars, decreasing = TRUE))[1:50]
heatmap_data <- log_expr[top_features, ]

# Etiquetas de columnas
sample_annots <- as.data.frame(colData(se)[, c("Treatment")])

# Asegurarse de que la anotación sea válida
sample_annots <- as.data.frame(colData(se)[, "Treatment", drop = FALSE])

# Convertir a factor si no lo es
sample_annots$Treatment <- as.factor(sample_annots$Treatment)

# Verifica estructura
str(sample_annots)
```

```
## 'data.frame':    54 obs. of  1 variable:
##  $ Treatment: Factor w/ 6 levels "DOG0","DOG1",..: 1 1 1 1 1 1 1 1 1 2 ...
```

```r
# Heatmap
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.4.3
```

```r
pheatmap(heatmap_data,
         scale = "row",
         annotation_col = sample_annots,
         show_rownames = FALSE,
         clustering_distance_rows = "euclidean",
         clustering_method = "complete",
         main = "Heatmap de las 50 features más variables")
```

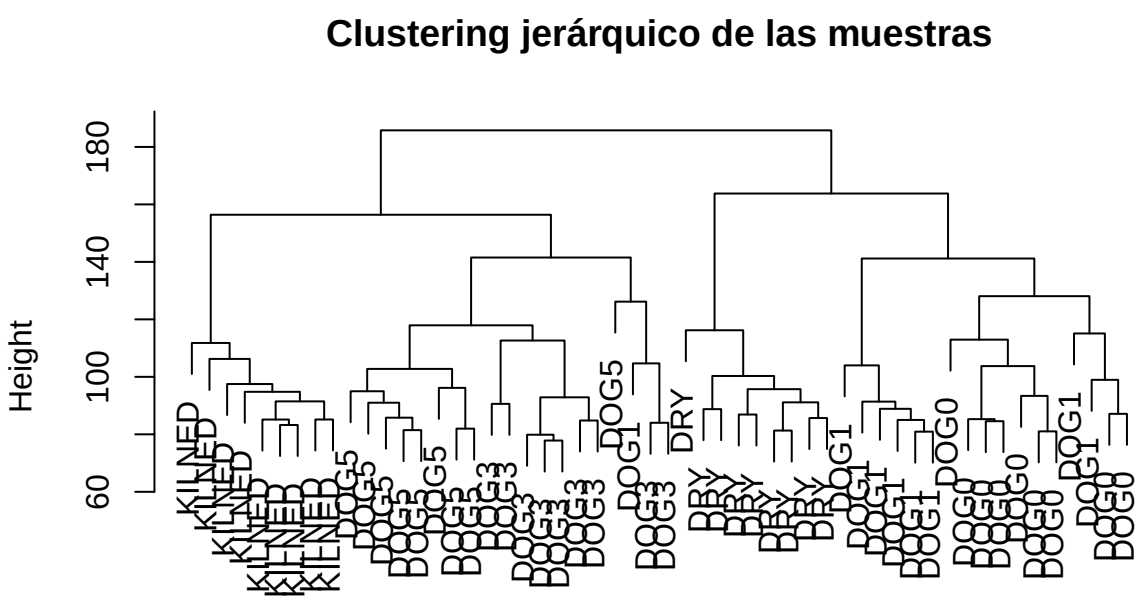


El heatmap revela una agrupación clara de las muestras según el tratamiento, con patrones de expresión diferenciados en las 50 features más variables. Cada fila representa una feature metabolómica identificada por su combinación de m/z y tiempo de retención, ordenada en el eje Y según la similitud de su patrón de abundancia entre muestras. Las muestras de un mismo grupo tienden a agruparse juntas, lo que indica coherencia interna, mientras que los tratamientos KILNED y DRY presentan perfiles notablemente distintos respecto al resto. Se observa una región central de features con alta intensidad relativa en DOG3 y DOG5, que contrasta con la baja intensidad en DOG0 y DRY, lo que sugiere que estas features podrían estar relacionadas con las condiciones experimentales aplicadas.

Clustering jerárquico de muestras

Seguidamente, creamos un dendrograma por clustering para ver agrupamientos por perfil metabolómico:

```
# Distancias entre muestras (usamos datos log-transformados y escalados)
dist_samples <- dist(t(log_expr_scaled), method = "euclidean")

# Clustering
hc <- hclust(dist_samples, method = "complete")

# Dendrograma
plot(hc, labels = colData(se)$Treatment,
     main = "Clustering jerárquico de las muestras",
     xlab = "", sub = "")
```



El clustering ha agrupado las muestras según la similitud de sus perfiles metabolómicos. Se observa una separación clara entre grupos de tratamiento, con KILNED y DRY formando ramas independientes. Los tratamientos DOG0, DOG1, DOG3 y DOG5 se agrupan en ramas más próximas, aunque con subdivisiones que reflejan diferencias internas. La organización general del dendrograma indica que el tratamiento tiene un efecto importante en la composición metabólica de las muestras, y que las réplicas dentro de cada grupo presentan un alto grado de similitud.

Interpretación de resultados

El análisis exploratorio del perfil metabolómico mediante LC-MS permite identificar patrones diferenciados entre tratamientos aplicados durante el proceso de malteado. La reducción de dimensionalidad mediante PCA, el mapa de calor de las variables más variables y el clustering jerárquico de muestras muestran una agrupación coherente por etapa, con especial diferenciación en las fases de kilning y germinación avanzada. Estos resultados sugieren que los tratamientos aplicados inducen cambios consistentes en el metabolismo de la cebada, y que dichos cambios pueden ser aprovechados como base para el desarrollo de indicadores de calidad o el seguimiento del proceso.

Referencias

Rani, H., & Whitcomb, S. J. (2025). Integrative LC-MS and GC-MS metabolic profiling unveils dynamic changes during barley malting. Food Chemistry, 463https://10.1016/j.foodchem.2024.141480