PEC1-ML-3

Diego Nuevo Sánchez

Contents

Resumen

Se ha creado un objeto SummarizedExperiment utilizando la librería POMA de bioconductor. Esto nos permite realizar estudios de manera más eficiente, como podría ser el estudio de PAC (Análisis de Componentes Principales), PLS o un ANOVA. No se aprecian diferencias significativas entre los distintos grupos con los análisis realizados.

Objetivos

El objetivo principal es estudiar cómo la remodelación de fosfatidilinositol mediada por la lisofosfolípido aciltransferasa 11 (MBOAT7) influye en la composición de especies de fosfatidilinositol (PI) en el córtex cerebral de ratones embrionarios KO y los WT para MBOAT7.

Diferencias entre las clases ExpressionSet y SummarizedExperiment

Ambas son estructuras de datos en R utilizadas para almacenar y manejar datos experimentales en bioinformática:

- ExpressionSet se utiliza principalmente para ensayos con microarrays, donde tienes una única matriz de datos de expresión, donde las filas representan las características y las columnas, las muestras. No incluye información sobre rangos genómicos de manera nativa.
- SummarizedExperiment se utiliza principalmente para los datos de secuenciación de nueva generación (NGS), permitiendo integrar múltiples matrices de datos y metadatos (assays) de manera más eficiente. Ofrece soporte directo para rangos genómicos a través de la clase “GRanges”.

Métodos

Se ha utilizado la metodología descrita en POMA workflow de bioconductor, información facilitada en la actividad 1.3.

Elegí el estudio "Phosphatidylinositol remodeling by lysophospholipid acyltransferase 11 ensures embryonic radial glial cell integrity", de identificador ST003747_AN006153 en metabolomic workbench.

Para este trabajo se ha utilizado el archivo mwTab en texto porque es más sencillo de procesar. Los datos más importantes en el fichero son "Sample_data" y "Metabolite_data", donde Sample_data contiene los metadatos de las muestras (como genotipo del ratón, edad del ratón), mientras que Metabolite_data contiene los datos de la concentración de los metabolitos en pmol/cortex.

```
# Cargamos la librería para lectura de archivos
library(readr)
library(tibble)
library(dplyr)
```

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

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

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

```
#Cargamos todas las líneas del archivo txt de metabolomicWorkbench
mwtab_lines <- readLines("ST003747_AN006153.txt")

#Identificamos el inicio y fin de nuestros datos de interés (en este caso
#Son los datos cuantitativos de los metabolitos)
inicio <- grep("#SUBJECT_SAMPLE_FACTORS:", mwtab_lines)
fin <- grep("#COLLECTION", mwtab_lines)

# Extraemos la sección que contiene la matriz de datos
data_section <- mwtab_lines[(inicio + 1):(fin - 1)]

# La convertimos en una tabla
sample_data <- read.table(text = data_section, header = FALSE, sep = "|", stringsAsFactors = FALSE, dec

sample_data <- sample_data[,-1]
sample_data <- sample_data[,-5]
sample_data<- sample_data[,-2]
sample_data <- sample_data %>%
  rename( Samples = V2, Gestational_age= V4, Genotipe= V5)
```

```
#repetimos el proceso con los metadatos de los metabolitos (mismo txt)
mwtab_lines <- readLines("ST003747_AN006153.txt")

inicio <- grep("MS_METABOLITE_DATA_START", mwtab_lines)
```

```
fin <- grep("MS_METABOLITE_DATA_END", mwtab_lines)

data_section <- mwtab_lines[(inicio+1):(fin - 1)]

metabolites <- read.table(text = data_section, header = TRUE, sep = "\t", stringsAsFactors = FALSE, dec
metabolites<- metabolites[-1,]
metabolites <- metabolites %>%
  rename(metabolites = Samples)
sample_data$metabolites<- metabolites$metabolites
sample_data <- sample_data[,c("metabolites", "Genotipe", "Samples", "Gestational_age")]
sample_data$Genotipe <- factor(sample_data$Genotipe,
                               levels = c(" Mboat7+/-"," Mboat7-/-"),
                               labels = c("WT", "KO"))

metabolites[,-1] <- lapply(metabolites[,-1], function(x) as.numeric(as.character(x)))
```

Una vez separados los datos de interés, podemos crear el objeto SummarizedExperiment:

```
library(POMA)
```

```
## Welcome to POMA!
## Version 1.16.0
## POMAShiny app: https://github.com/pcastellanoescuder/POMAShiny
```

```
poma_obj <- PomaCreateObject(metadata = sample_data, features = metabolites)
head(poma_obj)
```

```
## class: SummarizedExperiment
## dim: 6 26
## metadata(0):
## assays(1): ''
## rownames(6): metabolites E11.5_Het_1_PS_PI ... E11.5_Het_4_PS_PI
##   E11.5_KO_1_PS_PI
## rowData names(0):
## colnames(26): PI 32:1 PI 34:0 ... PS 40:5 PS 42:5
## colData names(3): Genotipe Samples Gestational_age
```

Una vez creado nuestro SummarizedExperiment, el primer paso es depurarlo de valores faltantes o nulos (en el caso de que hubiera):

```
#limpiamos los valores NA de nuestros datos
imputed <- poma_obj %>%
  PomaImpute(method = "knn", zeros_as_na = TRUE, remove_na = TRUE, cutoff = 20)
```

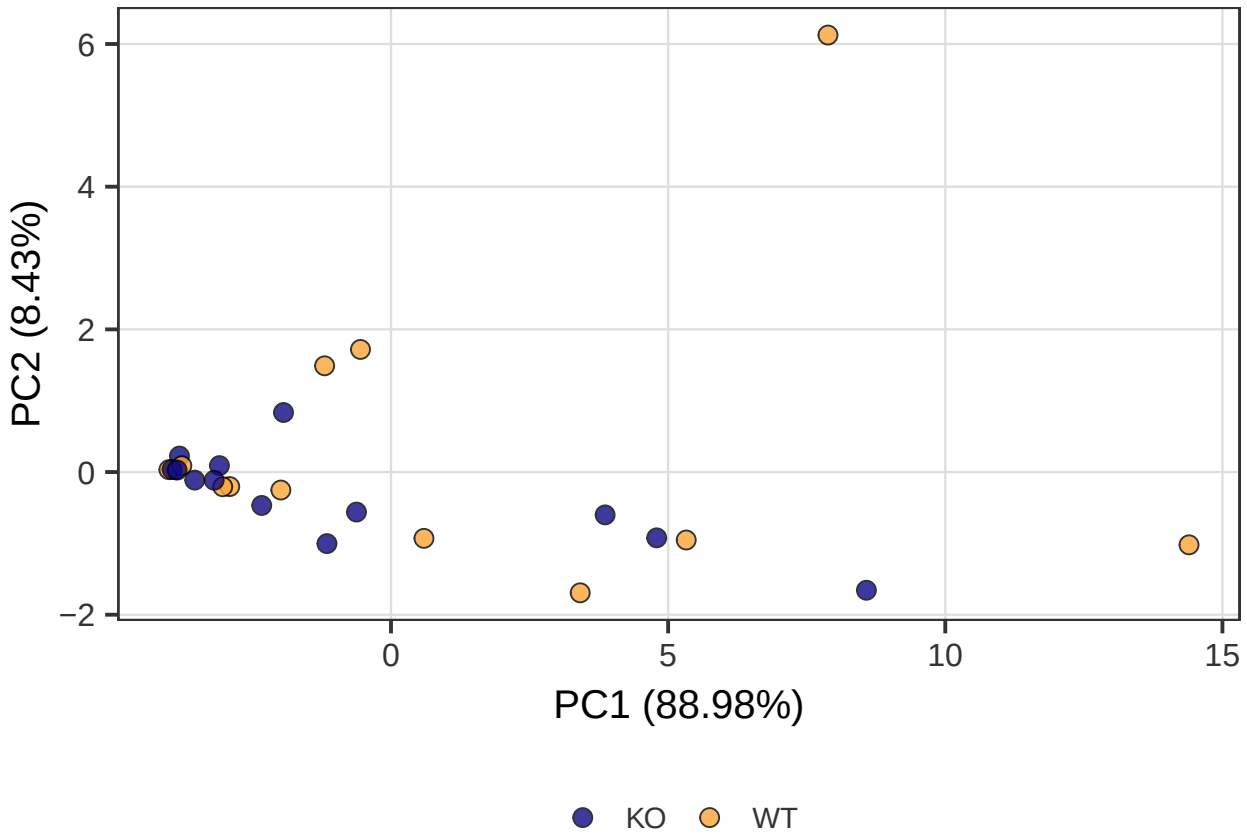
```
## 1 features removed.
```

```
normalized <- imputed %>%
  PomaNorm(method = "log_pareto")
```

Resultados

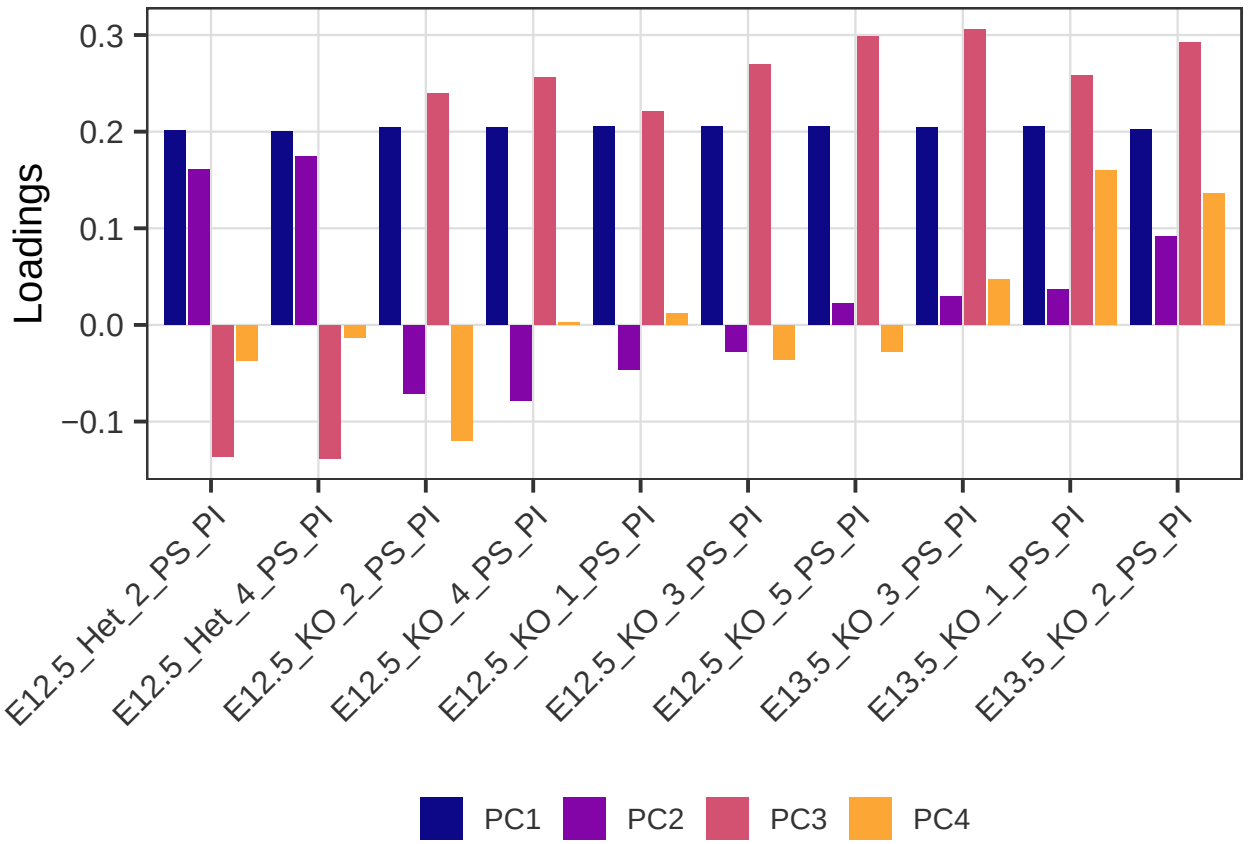
Una vez depurado y filtrado nuestro SE (SummarizedExperiment), podemos hacer un análisis de los componentes principales (PCA). No hace falta normalizar previamente los datos porque la función PomaPCA lo hace automáticamente.

```
PCA <- PomaPCA(imputed)
print(PCA$factors_plot)
```



PC1 explica un 89% de la varianza, por lo que al agruparse principalmente a la izquierda los ratones, se puede asumir que existe significancia estadística. Además, hay claras diferencias entre los ratones WT y KO.

```
print(PCA$loadings_plot)
```

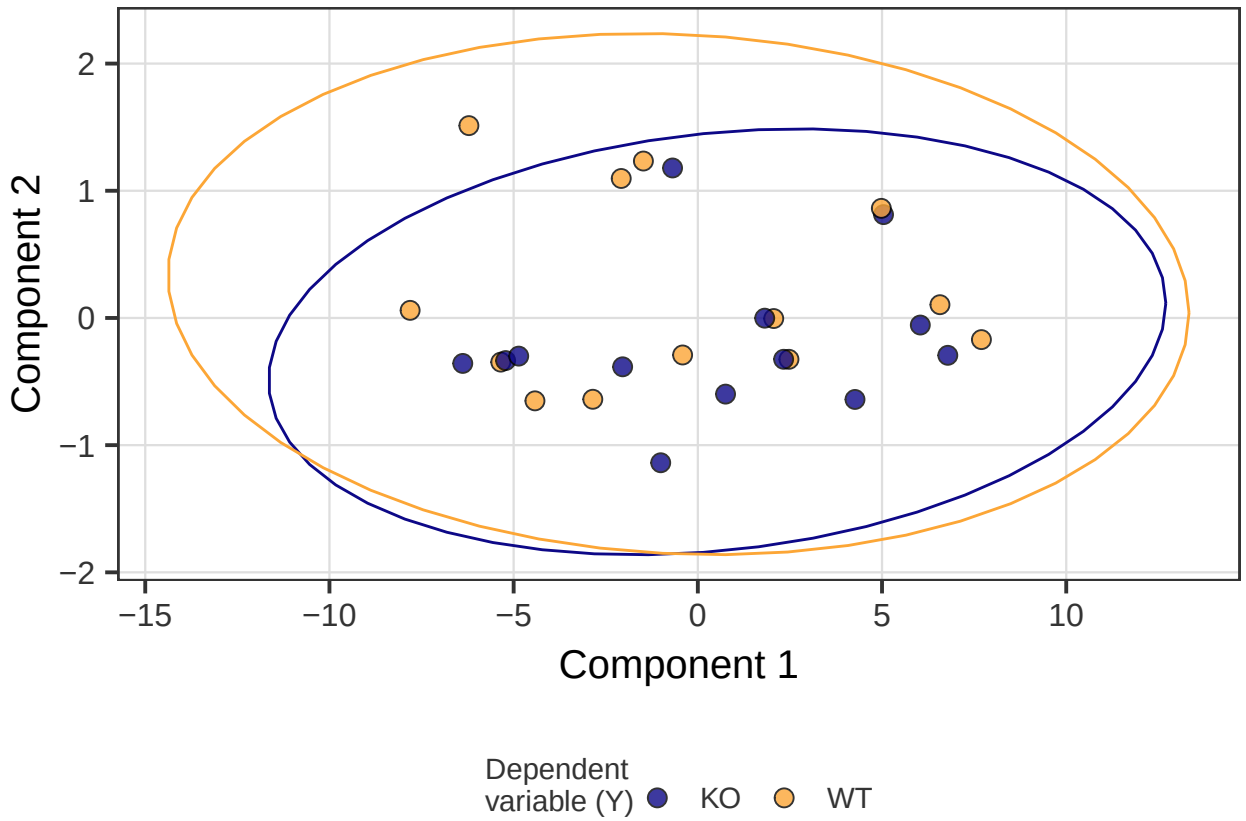


Este segundo gráfico representa las contribuciones más importantes de algunas muestras para explicar cada PC.

Observamos valores fuertes negativos en PC2 que podrían reflejar una diferencia entre KO y WT (Het). PC3 y PC4 son las de mayor variabilidad pero hemos visto que tienen poca significancia en este caso.

Por último, realizamos un análisis PLS para ver la separación por grupos.

```
plsda_result <- PomaPLS(
  data = normalized,
  method = "plsda",  # Para análisis discriminante (clases)
  ncomp = 3          # "ncomp" en lugar de "components"
)
plsda_result$factors_plot
```



Los datos están bastante dispersos alrededor del componente 1, lo que sugiere diferencias biológicas entre KO y WT.

Para salir de dudas, vamos a realizar un ANOVA:

```
resultados <- PomaUnivariate(
  data = imputed,
  method = "anova",
  covs = NULL,          # Sin covariables
  adjust = "fdr"        # Corrección por FDR
)
resultados_principales <- resultados$result
resultado_mas_significativo <- resultados_principales %>%
  dplyr::arrange(adj_pvalue) %>%
  dplyr::slice(1)
resultado_mas_significativo
```

```
## # A tibble: 1 x 7
##   feature           pvalue adj_pvalue mean_KO mean_WT sd_KO sd_WT
##   <chr>              <dbl>      <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 E12.5_Het_1_PS_PI  0.220      0.705    402.    751.  498.  867.
```

Como podemos observar (he puesto solo el p-valor más bajo para no poner la tabla entera), al realizar un ANOVA, salimos de dudas. No se observan diferencias estadísticamente significativas entre los diferentes grupos.

Discusión

Puede que utilizar un SummarizedExperiment no sea el mejor enfoque para este dataset, la forma en la que están dispuestos los datos en este caso específico hace bastante complicado manejarlos porque se opacan de primeras algunos datos como podría ser el tipo de metabolito PI, haciendo que sea complicado observar esas diferencias en los gráficos. De todos modos la posibilidad de tener todos los datos organizados de manera sencilla y automática puede ayudar mucho a la hora de realizar análisis de datos. He notado que a la hora de las preferencias de elegir nombres de los metadatos, a veces elige de manera automática las primeras columnas, como por ejemplo en los gráficos que se generan al hacer el PCA o al intentar hacer boxplot (se genera con las dos primeras columnas para los nombres).

Existen otros problemas con la asignación automática, como por ejemplo que al hacer un PomaCorr (correlaciones desde la librería POMA) me asigna automáticamente los mismos parámetros para calcular la correlación.

Conclusiones

A priori no se observan grandes diferencias en el perfil lipídico entre los ratones con MBOAT7 WT y KO, tal y como se ve en el anova. Se requieren más análisis para terminar de alcanzar los resultados observados en el experimento.

Referencias

https://github.com/Newrg/PEC1.git