Fernandez-Blanco-Concepcion-PEC1

Coqui Fernandez Blanco

2024-10-30

PEC 1: Introducción a los datos ómicos

1. Introducción

En este informe se analiza un dataset de fosfoproteómica para explorar la diferenciación entre dos grupos tumorales: MSS y PD. El objetivo es identificar fosfopéptidos que permitan diferenciar estos grupos utilizando análisis estadístico y visualización.

Este análisis incluye la selección y preparación de los datos, la creación de un contenedor `SummarizedExperiment`, la exploración de datos, y un análisis multivariante mediante Componentes Principales (PCA).

En mi caso, he seleccionado el siguiente dataset: **Datasets/2018-Phosphoproteomics**

Detalles del dataset seleccionado:

El conjunto de datos adjunto se ha obtenido a partir de un experimento de fosfoproteómica que se llevó a cabo para analizar (3 + 3) modelos PDX de dos subtipos diferentes utilizando muestras enriquecidas en fosfopéptidos. Se realizó un análisis de LC-MS con 2 réplicas técnicas en cada muestra. El conjunto de resultados consistió en abundancias normalizadas de señales de MS para aproximadamente 1400 fosfopéptidos.

Objetivo del análisis: Buscar fosfopéptidos que permitan diferenciar los dos grupos tumorales. Esto debe hacerse tanto con análisis estadístico como con visualización. Los datos se han proporcionado en un archivo de Excel: TIO2+PTYR-human-MSS+MSIvsPD.XLSX.

Los grupos se definen como:

- Grupo MSS: Muestras M1, M5 y T49.

- Grupo PD: Muestras M42, M43 y M64, con dos réplicas técnicas para cada muestra.

2. Preparación y Creación del Contenedor SummarizedExperiment

Instalación de Paquetes y Carga de Librerías

```
#instalamos los paquetes necesarios
#install.packages("BiocManager")
#BiocManager::install("SummarizedExperiment")
#install.packages("readxl")
library(SummarizedExperiment)
```

```
## Warning: package 'SummarizedExperiment' was built under R version 4.3.1
```

```
## Loading required package: MatrixGenerics
```

```
## Warning: package 'MatrixGenerics' was built under R version 4.3.1
```

```
## Loading required package: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 4.3.3
```

```
##
## Attaching package: '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
```

```
## Loading required package: GenomicRanges
```

```
## Warning: package 'GenomicRanges' was built under R version 4.3.1
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Warning: package 'BiocGenerics' was built under R version 4.3.1
```

```
##
## Attaching package: '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, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Warning: package 'S4Vectors' was built under R version 4.3.2

##
## Attaching package: 'S4Vectors'

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

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

## Loading required package: IRanges

## Warning: package 'IRanges' was built under R version 4.3.1

##
## Attaching package: 'IRanges'

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

## Loading required package: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.3.3

## Loading required package: Biobase

## Warning: package 'Biobase' was built under R version 4.3.1

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
```

```
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

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

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

library(readxl)

## Warning: package 'readxl' was built under R version 4.3.3

library(Biobase)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: '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
```

Lectura de los datos e identificación de abundancias

```
datos <- read_excel("C:/Users/fernac53/Documents/1.1._Curso_ Bioinf&Bioest/Asignaturas/Datos_omicos/TIO2+PTYR-human-MSS+MSIvsPD.XLSX",sheet=1)
targets <- read_excel(path= "C:/Users/fernac53/Documents/1.1._Curso_ Bioinf&Bioest/Asignaturas/Datos_omicos/TIO2+PTYR-human-MSS+MSIvsPD.XLSX", sheet=2)

## New names:
## • `Sample` -> `Sample...1`
## • `Sample` -> `Sample...2`

dim(datos)

## [1] 1438   18

colnames(datos)

##  [1] "SequenceModifications" "Accession"             "Description"
##  [4] "Score"                 "M1_1_MSS"              "M1_2_MSS"
##  [7] "M5_1_MSS"              "M5_2_MSS"              "T49_1_MSS"
## [10] "T49_2_MSS"             "M42_1_PD"              "M42_2_PD"
## [13] "M43_1_PD"              "M43_2_PD"              "M64_1_PD"
## [16] "M64_2_PD"              "CLASS"                 "PHOSPHO"

abundance_data <- as.matrix(datos[, 5:16])  #abundancia
rownames(abundance_data) <- datos$SequenceModifications  #asignar nombres de fila

#vector de grupos
groups <- c(rep("MSS", 6), rep("PD", 6))

#dataframe de metadatos
metadata <- data.frame(Sample = colnames(abundance_data), Group = groups, Phenotype = targets$Phenotype) #con los datos de 'targets', la segunda hoja

#SummarizedExperiment con la nueva metadata
se <- SummarizedExperiment(assays = list(counts = abundance_data), colData = metadata)
se
```

```
## class: SummarizedExperiment
## dim: 1438 12
## metadata(0):
## assays(1): counts
## rownames(1438): LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation
##   VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho ...
##   YQDEVFGGFVTEPQEESEEEVEEPEER[17] Phospho YSPSQNSPIHHIPSRR[1]
##   Phospho|[7] Phospho
## rowData names(0):
## colnames(12): M1_1_MSS M1_2_MSS ... M64_1_PD M64_2_PD
## colData names(3): Sample Group Phenotype

#Guardamos
write.csv(abundance_data, "abundance_data.csv", row.names = TRUE)
save(se, file = "SummarizedExperiment_data.Rda")
```

Este contenedor nos permite almacenar tanto los datos de abundancia como los metadatos de las muestras, facilitando un análisis estructurado y accesible.

3. Exploración de los Datos

```
summary(assay(se, "counts"))
```

```
##     M1_1_MSS          M1_2_MSS          M5_1_MSS          M5_2_MSS      
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:    5653   1st Qu.:    5497   1st Qu.:    2573   1st Qu.:    3273
##  Median :   30682   Median :   26980   Median :   20801   Median :   26241
##  Mean   :  229841   Mean   :  253151   Mean   :  232967   Mean   :  261067
##  3rd Qu.:  117373   3rd Qu.:  113004   3rd Qu.:  113958   3rd Qu.:  130132
##  Max.   :16719906   Max.   :43928481   Max.   :15135169   Max.   :19631820
##    T49_1_MSS         T49_2_MSS          M42_1_PD          M42_2_PD      
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:    9306   1st Qu.:    8611   1st Qu.:    5341   1st Qu.:    4216
##  Median :   55641   Median :   46110   Median :   36854   Median :   30533
##  Mean   :  542449   Mean   :  462616   Mean   :  388424   Mean   :  333587
##  3rd Qu.:  223103   3rd Qu.:  189141   3rd Qu.:  180252   3rd Qu.:  152088
##  Max.   :49218872   Max.   :29240206   Max.   :48177680   Max.   :42558111
##     M43_1_PD          M43_2_PD          M64_1_PD          M64_2_PD      
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:   19641   1st Qu.:   17299   1st Qu.:   11038   1st Qu.:    8660
##  Median :   67945   Median :   59607   Median :   52249   Median :   47330
##  Mean   :  349020   Mean   :  358822   Mean   :  470655   Mean   :  484712
##  3rd Qu.:  205471   3rd Qu.:  201924   3rd Qu.:  209896   3rd Qu.:  206036
##  Max.   :35049402   Max.   :63082982   Max.   :71750330   Max.   :88912734
```

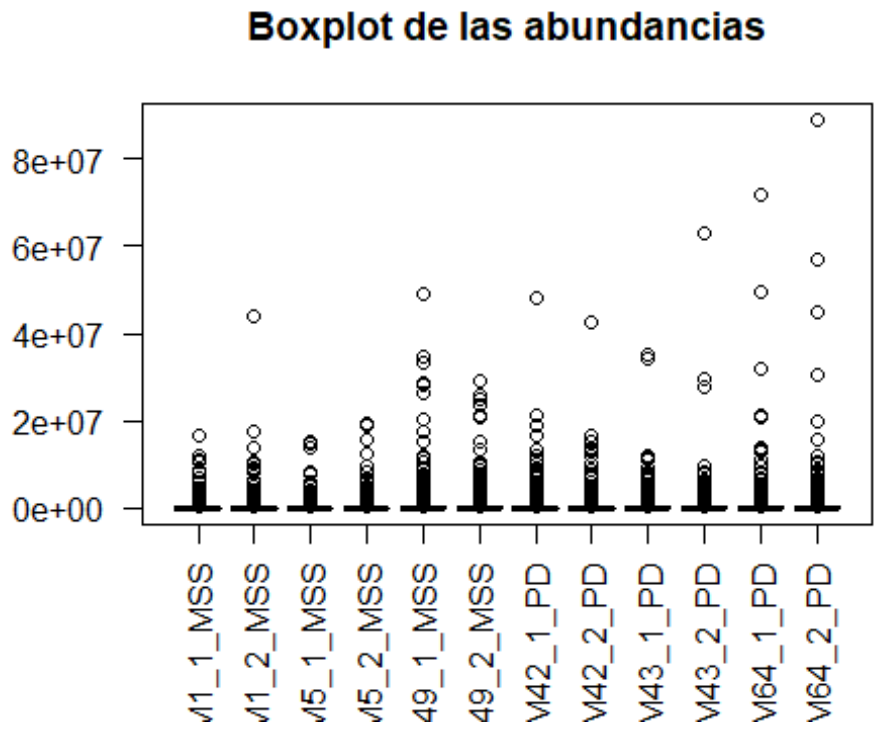
Los datos de abundancia de fosfopeptidos muestran una variabilidad significativa, con algunos valores extremadamente altos que podrían influir en las medias y análisis estadísticos. Las muestras del grupo PD presentan mayores abundancias promedio en comparación con el grupo MSS, lo que sugiere diferencias relevantes entre los grupos.

Visualización de los datos

Boxplot

Añadimos visualización de los datos proporcionados

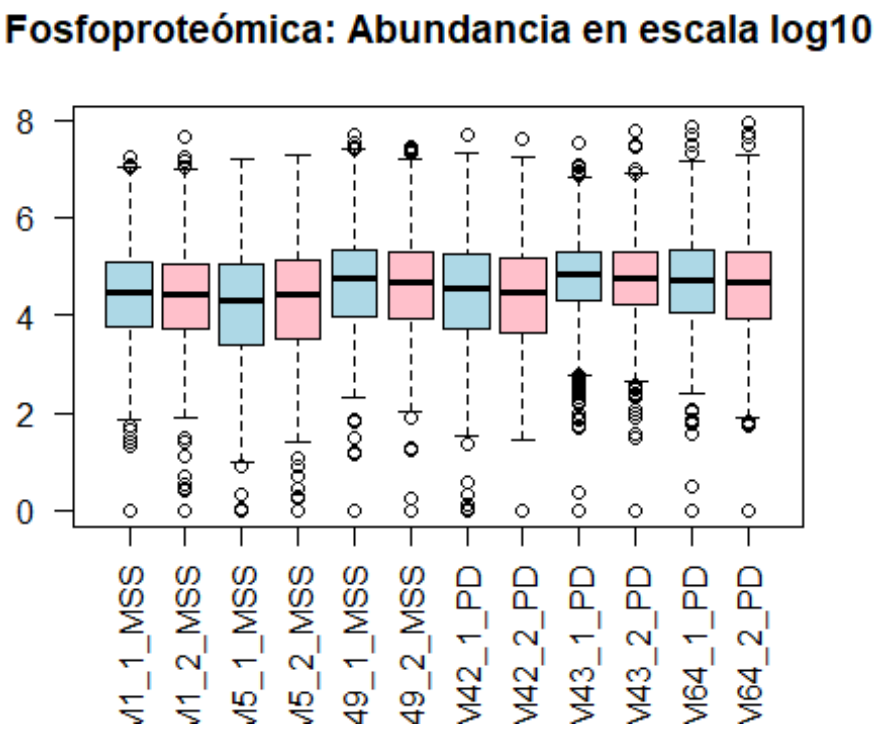
```
boxplot(assay(se, "counts"), main = "Boxplot de las abundancias", las = 2, co
l = c("lightgreen", "lightcoral"))
```



Debido a la naturaleza de los datos (la escala en la que están las abundancias), el gráfico no sale bien, por lo que sería mucho más útil visualizar los datos en escala logarítmica:

Boxplot arreglado

```
boxplot(log10(assay(se, "counts") + 1), las = 2,
        main = "Fosfoproteómica: Abundancia en escala log10", col = c("lightb
lue", "pink"))
```



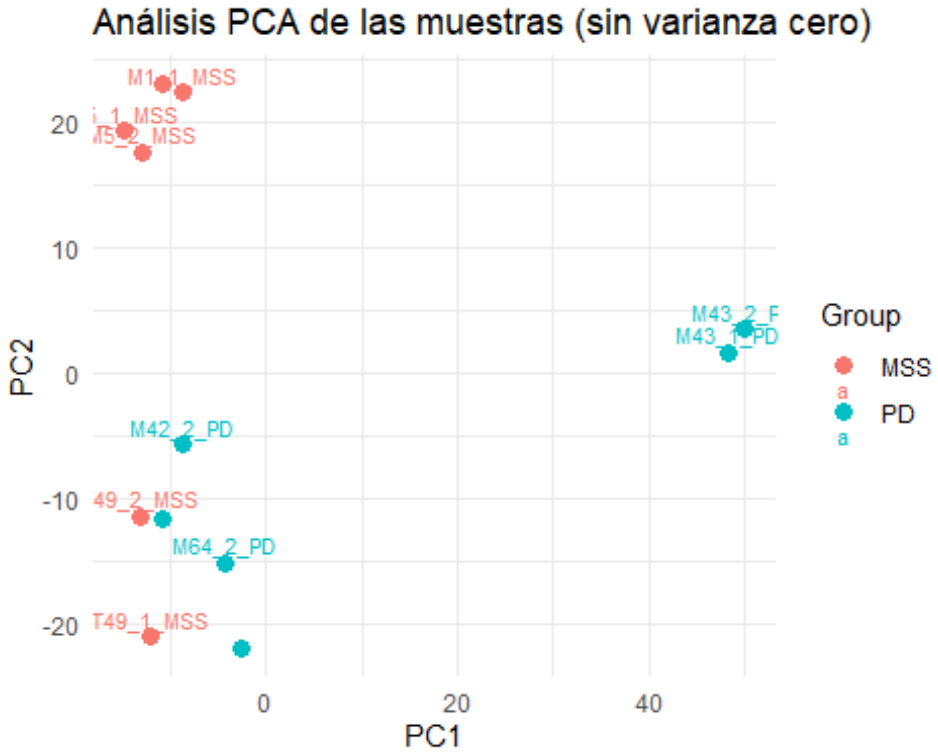
Análisis de Componentes Principales (PCA)

Para realizar el PCA, hemos usado los datos de se, pero inicialmente, encontramos un error debido a que alguna de las muestras tiene una varianza igual a 0: Error in prcomp.default(t(assay(se, “counts”)), scale. = TRUE) : cannot rescale a constant/zero column to unit variance. Por lo que hemos filtrado el set apra realizar el PCA:

```
row_variance <- apply(assay(se, "counts"), 1, var, na.rm = TRUE)
#como da error sin filtrarm hemos de filtrar filas con varianza mayor que cer
o
se_filtered <- se[row_variance > 0, ]
se_filtered
```

```
## class: SummarizedExperiment
## dim: 1436 12
## metadata(0):
## assays(1): counts
## rownames(1436): LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation
##   VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho ...
##   YQDEVFGGFVTEPQEESEEEVEEPEER[17] Phospho YSPSQNSPIHHIPSRR[1]
##   Phospho|[7] Phospho
## rowData names(0):
## colnames(12): M1_1_MSS M1_2_MSS ... M64_1_PD M64_2_PD
## colData names(3): Sample Group Phenotype
```

```r
pca_res <- prcomp(t(assay(se_filtered, "counts")), scale. = TRUE)
pca_data <- data.frame(pca_res$x, Group = colData(se_filtered)$Group)
pca_data$Sample <- rownames(pca_data)
ggplot(pca_data, aes(x = PC1, y = PC2, color = Group)) +
  geom_point(size = 3) +
  geom_text(aes(label = Sample), vjust = -0.5, hjust = 0.5, size = 3, check_o
verlap = TRUE) +
  labs(title = "Análisis PCA de las muestras (sin varianza cero)", x = "PC1",
y = "PC2") +
  theme_minimal()
```



En el gráfico de componentes principales puede verse bien la distribución de las muestras en el espacio de los dos primeros componentes principales (PC1 y PC2). Se observa una clara separación entre los grupos MSS y PD, lo que sugiere que hay diferencias significativas en las abundancias de los fosfopéptidos que permiten distinguir los dos grupos tumorales. Las muestras del grupo MSS se agrupan en una región diferente a las del grupo PD, lo que refuerza la idea de que los datos tienen características que permiten la diferenciación entre los subtipos estudiados.También parece haber mucha más variabilidad dentro del grupo PD.

Conclusiones

Este análisis de fosfoproteómica muestra una diferenciación entre los grupos tumorales MSS y PD. Los resultados del PCA indican una separación significativa en las abundancias, lo que sugiere que ciertos fosfopéptidos podrían ser biomarcadores útiles para diferenciar entre estos subtipos tumorales.

Posibles limitaciones y mejoras

- Limitaciones: La alta variabilidad en las abundancias podría afectar la robustez de algunos análisis estadísticos.
- Mejoras: Continuar con anäalisis estadäistico en profundidad para poder comprobar diferenciación entre dos grupos tumorales.

Repositorio en github

El contenido de la PEC1 se encuentra en el siguiente repositorio:

https://github.com/cfernandezblan/FERNANDEZ-Blanco-Concepcion-PEC1