# PEC1_ADO

Francisco A. Aparicio González



# Abstract

*En este informe analizamos un conjunto de datos de fosfoproteómica obtenidos mediante espectrometría de masas (LC-MS) en modelos tumorales humanos tipo PDX. El objetivo principal ha sido detectar diferencias en los perfiles de fosforilación entre dos subtipos tumorales: MSS y PD.*

*Tras importar y preparar los datos, se ha llevado a cabo una exploración inicial de la calidad, seguida de un análisis estadístico usando el paquete* `limma`*. Finalmente, se visualizaron los resultados con técnicas como PCA, clustering y heatmaps. Los resultados muestran diferencias claras entre ambos grupos, y permiten identificar un conjunto de fosfopéptidos potencialmente relevantes a nivel biológico.*

*Todo el contenido se subió al siguyinte repositorio público de GitHub:*
*https://github.com/fadh13/Aparicio-Gonzalez-Francisco-PEC1*

# Objetivos

- Comparar los perfiles de fosforilación entre los grupos MSS y PD
- Detectar fosfopéptidos diferencialmente abundantes entre ambos subtipos tumorales.
- Evaluar la calidad de los datos y la consistencia entre réplicas.
- Visualizar los resultados mediante técnicas gráficas para interpretar posibles agrupamientos.
- Preparar los datos y resultados para su publicación en un repositorio abierto (GitHub).

# Materiales y métodos

Los datos utilizados provienen del repositorio metaboData, concretamente del dataset **2018-Phosphoproteomics**. Este contiene abundancias normalizadas de fosfopéptidos obtenidas por LC-MS a partir de modelos PDX, en los que se comparan dos subtipos tumorales distintos: **MSS** y **PD**. Cada muestra cuenta con dos réplicas técnicas, y en total se incluyen 12 columnas correspondientes a seis pacientes (tres por grupo) y más de 1400 fosfopéptidos.

El análisis se ha desarrollado en R, siguiendo los siguientes pasos:

1. **Importación y preprocesado**:
    a. Lectura del archivo Excel con los datos de abundancia.
    b. Filtrado de las columnas necesarias y transformación a una matriz numérica
    c. Creación de un objeto `SummarizedExperiment` que incluye tanto la matriz como los metadatos (grupo y réplica).
2. **Análisis exploratorio**:
    a. Visualización de histogramas, boxplots y gráficos de correlación entre réplicas.
    b. PCA para observar separación entre grupos.
    c. Clustering jerárquico para confirmar la coherencia del agrupamiento.
3. **Análisis estadístico**:
    a. Se utilizó el paquete `limma` para identificar fosfopéptidos diferencialmente abundantes entre MSS y PD.
    b. Se aplicó un modelo lineal, moderación bayesiana y corrección de p-valores (FDR).
    c. Los resultados se representaron mediante volcano plot y heatmap.
4. **Exportación y documentación**:
    a. Se guardaron los datos procesados y el objeto `SummarizedExperiment` en archivos `.csv` y `.Rda.` También se generó un `README.md`
    b. Se subieron todos los archivos al repositorio de GitHub

# Resultados

## Introducción

Ya hemos descargado los datos desde el repositorio de GitHub y los hemos organizado en nuestro directorio de trabajo. Para este análisis, utilizaremos el dataset “2018-Phosphoproteomics”, que proviene de un experimento de fosfoproteómica realizado

sobre modelos PDX (3 + 3) de dos subtipos tumorales distintos. Las muestras, enriquecidas en fosfopéptidos, fueron analizadas mediante LC-MS con dos réplicas técnicas por muestra.

El archivo de datos proporcionado es un Excel titulado *"TIO2+PTYR-human-MSS+MSIvsPD.XLSX"* e incluye abundancias normalizadas de aproximadamente 1400 fosfopéptidos. El objetivo del análisis es identificar fosfopéptidos capaces de diferenciar entre los dos grupos tumorales, combinando análisis estadístico y visualización. Los grupos se definen de la siguiente manera:

- **MSS**: Muestras M1, M5 y T4
- **PD**: Muestras M42, M43 y M64

La primera columna del archivo, SequenceModification, contiene los valores de abundancia para cada fosfopéptido y será la base para nuestro análisis. Las demás columnas pueden ser descartadas.

## Preprocesado y creación del objeto SummarizedExperiment

```
# Cargamos las librerías necesarias para el análisis

library(SummarizedExperiment)
library(readr)
library(Biobase)
library(readxl)
library(limma)
library(pheatmap)
```

### Carga del archivo Excel

```
# Indicamos la ruta del archivo descargado
 file_path <- "C:/Users/Fran/Desktop/Bioinformática/Análisis
Ómicos/data_ado_pec1/TIO2+PTYR-human-MSS+MSIvsPD.XLSX"

# Leemos el archivo Excel y guardamos el contenido en un objeto llamado data_0
 data_0 <- read_excel(file_path)
# Visualización general del contenido
 summary(data_0)
```

### Selección de columnas relevantes

Filtramos solo las columnas que contienen información útil para el análisis. Nos interesa la columna de identificación y las columnas con las abundancias por muestra y réplica, infromación que extraímos del read.me.

```
# Nos quedamos con la columna de identificación y las columnas de abundancias
 ## La columna "SequenceModifications" servirá como identificador de fila
 ## Las columnas "M" contienen las abundancias para cada muestra y réplica técnica
data_1 <- data_0[, c("SequenceModifications",
```

```
"M1_1_MSS", "M1_2_MSS",
"M5_1_MSS", "M5_2_MSS",
"T49_1_MSS", "T49_2_MSS",
"M42_1_PD", "M42_2_PD",
"M43_1_PD", "M43_2_PD",
"M64_1_PD", "M64_2_PD")]
```

Comprobamos si hay identificadores de fosfopéptidos duplicados.

```
sum(duplicated(data_1$SequenceModifications))

## [1] 1
```

### Conversión a matriz de expresión

Transformamos el data frame en una matriz numérica que pueda ser usada por Bioconductor, y usamos los nombres de los fosfopéptidos como nombres de fila.

```
# Usamos la columna de identificador como nombres de fila
 # Transformamos el tibble a data.frame antes de trabajar con rownames
 # Usamos make.unique() para resolver duplicados

 data_1 <- as.data.frame(data_1)
 rownames(data_1) <- make.unique(as.character(data_1$SequenceModifications))
 data_1 <- data_1[, -1] # Eliminamos la columna ya usada como rownames
 data_1 <- as.matrix(data_1) # Convertimos a matriz numérica
```

### Creación de metadatos (colData)

Creamos los metadatos que describen a cada muestra (columna). Indicamos a qué grupo pertenece (MSS o PD) y el número de réplica técnica (1 o 2).

```
# Obtenemos los nombres de las columnas (muestras)
 sample_names <- colnames(data_1)

# Extraemos el grupo a partir del nombre de muestra
 group <- sub(".*_(MSS|PD)$", "\\1", sample_names)

 # Extraemos el número de réplica desde el nombre
 rep <- as.numeric(sub(".*_(\\d)_.*", "\\1", sample_names))

# Construimos el data frame de metadatos
 col_data <- data.frame(
 group = group,
 replicate = rep,
 row.names = sample_names)
```

### Creación del objeto SummarizedExperiment

Este objeto SE será el contenedor principal para realizar los análisis correspondientes en los siguientes pasos.

```
# Creamos el objeto SummarizedExperiment con datos y metadatos
se <- SummarizedExperiment(assays = list(counts = data_1), # Matriz de expresión
colData = col_data) # Metadatos de las muestras
```

## Análisis exploratorio del dataset

Nuestro objetivo a continuación será identificar fosfopéptidos diferenciales entre los grupos tumorales MSS y PD. Para ello, es esencial realizar primero un análisis exploratorio que nos permita evaluar la calidad general de los datos, detectar posibles muestras atípicas y/o observar patrones.

### Distribución de abundancia de fosfopéptidos (Anexo 1)

Primero representamos la distribución de abundancias por muestra usando histogramas. Todas siguen un patrón bastante típico en este tipo de datos: la mayoría de fosfopéptidos tienen valores bajos.

No se aprecian diferencias llamativas entre muestras, lo que nos indica que hay coherencia entre réplicas técnicas y entre los grupos MSS y PD.

```
# Histograma de intensidades de fosfopéptidos por muestra
par(mfrow = c(3, 4)) # Mostrar 12 gráficos (3 filas, 4 columnas)
for (i in 1:ncol(data_1)) { hist(data_1[, i],
main = colnames(data_1)[i],
col = "skyblue",
breaks = 30)}
```

### Boxplot de abundancia normalizada por muestra (Anexo 2)

El boxplot nos permite ver cómo se reparten las abundancias en cada muestra y detectar si hay alguna que se salga demasiado del patrón. Las medianas están bastante alineadas entre todas, los datos parecen bien normalizados y sin sesgos claros. Aunque hay algunos valores extremos (outliers), es algo esperado en este tipo de datos, donde unos pocos péptidos podrían estar mucho más expresados que otros.

```
# Boxplots para comparar abundancias entre muestras
boxplot(data_1,
```

```
las = 2,
col = "lightgreen")
```

### Correlación entre réplicas técnicas (Anexo 3)

Para comprobar la consistencia de las réplicas entre sí, generamos gráficos de dispersión (pairs plots) por grupo. Como vemos, la correlación entre réplicas es alta, esto nos da confianza en que el experimento ha salido bien y las mediciones son reproducibles.

```
# Correlación entre réplicas técnicas del mismo paciente - Grupo MSS
pairs(data_1[, c("M1_1_MSS", "M1_2_MSS",
"M5_1_MSS", "M5_2_MSS",
"T49_1_MSS", "T49_2_MSS")])

# Correlación entre réplicas técnicas del mismo paciente - Grupo PD
pairs(data_1[, c("M42_1_PD", "M42_2_PD",
"M43_1_PD", "M43_2_PD",
"M64_1_PD", "M64_2_PD")])
```

### Análisis de componentes principales (PCA) (Anexo 4)

A continuation relizamos un PCA para ver si los datos se separan bien entre grupos sin necesidad de etiquetarlos. El resultado es bastante claro: las muestras MSS (azul) y PD (rojo) aparecen agrupadas en zonas distintas del gráfico a pares, lo que indica que sus perfiles de fosforilación son distintos y consistentes dentro de cada grupo.

```
# Transponer la matriz: PCA requiere que las muestras estén en filas
data_t <- t(data_1)

# Eliminar fosfopéptidos con varianza cero para corregir errores en el PCA
data_t <- data_t[, apply(data_t, 2, var) != 0]

# Aplicamos PCA
pca <- prcomp(data_t, scale. = TRUE)

# Representamos los dos primeros componentes principales
plot(pca$x[, 1:2],
col = ifelse(col_data$group == "MSS", "blue", "red"),
pch = 19,
xlab = "PC1", ylab = "PC2")

legend("topright", legend = c("MSS", "PD"), col = c("blue", "red"), pch = 19)
```

### Clustering jerárquico de las muestras (Anexo 5)

El dendrograma muestra que las réplicas se agrupan entre sí, lo que confirma que el experimento ha sido consistente. y que las muestras MSS y PD también tienden a separarse en ramas distintas, como ya vimos en el PCA

```
# Calculamos distancias euclídeas entre columnas (muestras)
hc <- hclust(dist(t(data_1)))

# Mostramos el dendrograma resultante
plot(hc, main = "Clustering jerárquico de las muestras")
```

## Análisis estadístico para identificar fosfopéptidos diferenciales

En esta parte del análisis queremos ver qué fosfopéptidos están realmente diferenciando a los grupos MSS y PD. Para eso usamos el paquete limma, que nos permite ajustar un modelo lineal a los datos. Creamos una matriz de diseño para identificar a qué grupo pertenece cada muestra, y una matriz de contraste para compararlas directamente. Luego aplicamos una moderación bayesiana a las varianzas (algo muy útil cuando hay pocos replicados), y corregimos los p-valores con el método de Benjamini-Hochberg. Por último, visualizamos los resultados con un **volcano plot**, que nos deja ver de un vistazo qué péptidos son más significativos y están más cambiados, y un **heatmap** con los más relevantes para ver su patrón de expresión.

```
# Creamos el vector con los grupos biológicos
grupo <- factor(col_data$group, levels = c("MSS", "PD"))
# Creamos la matriz de diseño (modelo lineal)
matriz_d <- model.matrix(~0 + grupo)
colnames(design) <- levels(grupo)

# Ajustar el modelo lineal a los datos de abundancia
fit <- lmFit(data_1, matriz_d)
# Especificar la comparación entre grupos: PD - MSS
contraste <- makeContrasts(PDvsMSS = PD - MSS, levels = matriz_d)
# Aplicar el contraste al modelo ajustado
fit2 <- contrasts.fit(fit, contraste)
# Aplicar el modelo bayesiano moderado
fit2 <- eBayes(fit2)
```

```
## Warning: Zero sample variances detected, have been offset away from zero
```

```
# Obtener los resultados ordenados por significancia (ajuste BH)
resultados <- topTable(fit2, coef = "PDvsMSS", number = Inf, adjust = "BH")

# Contar cuántos fosfopéptidos presentan diferencias claras
sum(resultados$adj.P.Val < 0.05)
```

```
## [1] 109
```

### Volcano plot (Anexo 6)

El volcano plot nos ayuda a ver de forma rápida qué fosfopéptidos son significativos y cuánto cambian entre grupos. Los puntos rojos son los que tienen un P val < 0.05, y las líneas de referencia nos sirven para ubicar los umbrales de significancia y de cambio en la abundancia. Es una forma muy visual de detectar los candidatos más interesantes.

```
# Volcano plot
plot(resultados$logFC, -log10(resultados$adj.P.Val),
pch = 20,
col = ifelse(resultados$adj.P.Val < 0.05, "red", "grey"))
abline(h = -log10(0.05), col = "blue", lty = 2) # Umbral de significancia
abline(v = c(-1, 1), col = "darkgreen", lty = 2) # Umbral de cambio biológico
```

### Análisis con limma y visualización en heatmap

Se filtran los fosfopéptidos significativos y se representan en un mapa de calor.

```
# Filtrar los fosfopéptidos con p < 0.05
significativos <- resultados[which(resultados$adj.P.Val < 0.05), ]
pep_signif <- rownames(significativos)

# Extraer y escalar los datos para el heatmap
signif <- data_1[pep_signif, ]
escalado <- t(scale(t(signif)))

# Crear heatmap con anotación de grupo
pheatmap(escalado,
annotation_col = data.frame(Grupo = col_data$group),
annotation_colors = list(Grupo = c(MSS = "blue", PD = "red")),
show_rownames = FALSE)
```

### Heatmap de fosfopéptidos diferencialmente expresados (Anexo 7)

El heatmap muestra claramente la separación entre los grupos MSS y PD a partir de los fosfopéptidos más significativos. Las réplicas se agrupan como se espera de ellas, y se ven patrones de expresión bastante definidos, lo que apunta a que hay firmas moleculares consistentes detrás de esas diferencias.

```
# Listar los 10 primeros fosfopéptidos significativos
pep_signif <- rownames(resultados[resultados$adj.P.Val < 0.05, ])
head(pep_signif, 10)

## [1] "GVGYETILK[4] Phospho"
## [2] "AFGYYGPLR[4] Phospho"
## [3] "KASPEPPDSAEGALK[3] Phospho"
## [4] "KDPEDTGAEKSPTTSADLK[14] Phospho"
## [5] "IHDLEDDLEMSSDASDASGEEGGRVPK[10] Oxidation|[18] Phospho"
```

```
## [6] "LDQPVSAPPSPR[10] Phospho"
## [7] "PYQYPALTPEQK[4] Phospho"
## [8] "GIPLATGDTSPEPELLPGAPLPPPK[9] Phospho"
## [9] "HKAPGSADYGFAPAAGR[9] Phospho"
## [10] "EDSGTFSLGK[3] Phospho"
```

El análisis estadístico ha permitido identificar 109 fosfopéptidos con diferencias significativas (P valor < 0.05**)** entre los grupos MSS y PD. Estos péptidos podrían estar implicados en rutas de señalización celular diferencialmente activadas en cada tipo tumoral. En un estudio biomédico real, estos se mapearían a sus proteínas parentales y se analizarían mediante herramientas funcionales.

# Discusión y conclusiones

A lo largo del análisis hemos podido comprobar que los datos eran técnicamente consistentes: las réplicas técnicas se agrupan como deberían, y tanto el PCA como el clustering muestran una separación clara entre los grupos MSS y PD. Esto sugiere que hay diferencias reales en los perfiles de fosforilación entre ambos tipos tumorales. El uso del paquete limma nos permitió identificar más de un centenar de fosfopéptidos con diferencias estadísticamente significativas. Aunque no hemos profundizado en su interpretación, algunos de ellos podrían estar relacionados con procesos como la proliferación, la respuesta al estrés celular o rutas de señalización alteradas en ciertos tipos de cáncer utilizando bases de datos como KEGG. Este tipo de análisis es muy útil para generar hipótesis sobre posibles biomarcadores o mecanismos biológicos implicados en la progresión tumoral, aunque por supuesto haría falta validarlos experimentalmente.

Algunas conclusiones clave que podemos extraer son:

1. Hemos identificado diferencias claras en los perfiles de fosfopéptidos entre los grupos MSS y PD.
2. Los datos estaban bien normalizados y las réplicas técnicas mostraron una alta consistencia.
3. El análisis estadístico reveló 109 fosfopéptidos con diferencias significativas entre grupos.
4. Técnicas como PCA, clustering y heatmaps confirmaron la separación entre muestras y la validez de los resultados.
5. Este tipo de enfoque puede ser un punto de partida para estudios más profundos sobre funciones celulares y posibles biomarcadores.

# Preparación para GitHub

En esta sección se generan los archivos necesarios para documentar y compartir el proyecto a través de un repositorio en GitHub:

- Objeto de clase SummarizedExperiment en formato .Rda.
- Matriz de expresión en formato .csv.
- Metadatos de las muestras en .csv.
- Archivo README.md con la descripción del proyecto.
- Informe .pdf de 10 páginas más Anexo
- Archivo RMarkdown con el código ejecutable y el informe

### Guardar el objeto SummarizedExperiment

```
# Guardar el objeto 'se' (SummarizedExperiment)
save(se, file = "SummarizedExperiment_PEC1_fosfoproteomica.Rda")
```

### Exportar los datos y metadatos como archivos CSV

```
# Guardar la matriz de expresión
write.csv(data_1, file = "data_PEC1.csv", row.names = TRUE)

# Guardar los metadatos
write.csv(col_data, file = "colDataPEC1.csv", row.names = TRUE)
```

### Crear un archivo README

Hemos creado un archivo README directamente en el repositorio con la siguiente información:

*Primera práctica evaluable continua (PEC) de la asignatura Análisis de Datos Ómicos: Este repositorio contiene el código, los datos y el informe correspondiente al análisis de un dataset de fosfoproteómica comparando muestras de dos subtipos tumorales humanos (MSS y PD).*

*Informe final: Contenido:*

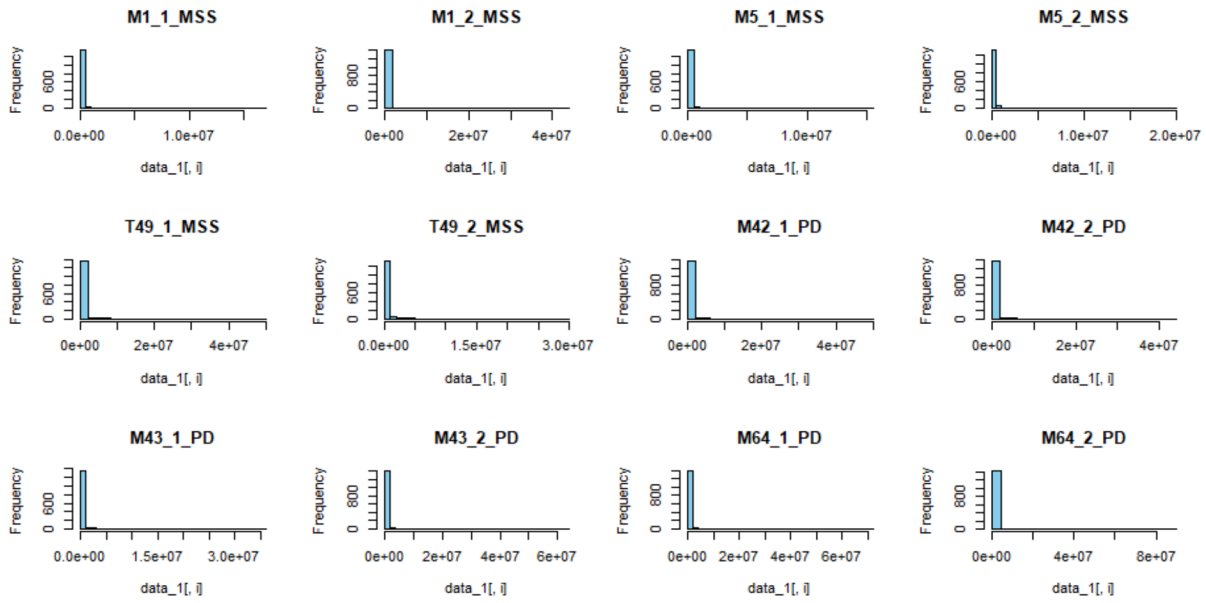
- *SummarizedExperiment_PEC1_fosfoproteomica: objeto con los datos y metadatos integrados*
- *data_PEC1.csv: matriz de abundancia normalizada de fosfopéptidos*
- *colDataPEC1.csv: metadatos de las muestras*
- *PEC1_ADO_Francisco_Aparicio: R Markdown con código completo en R y el informe final del análisis*
- *Informe final en archivo .pdf de 10 páginas más Anexo*

# Referencias

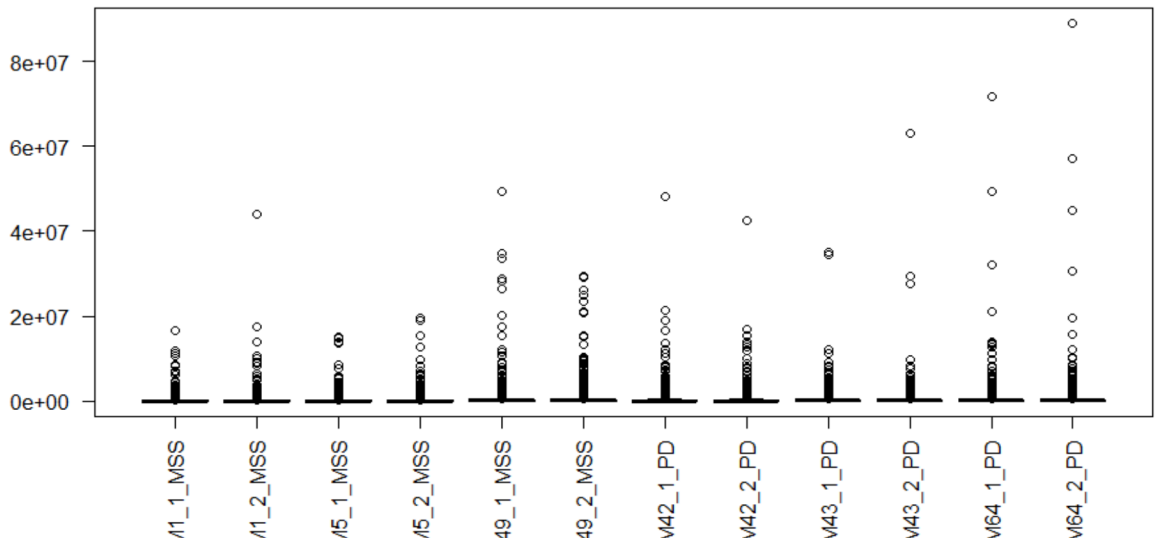

- Enalce al repositorio de Github: https://github.com/fadh13/Aparicio-Gonzalez-Francisco-PEC1
-

# Anexos

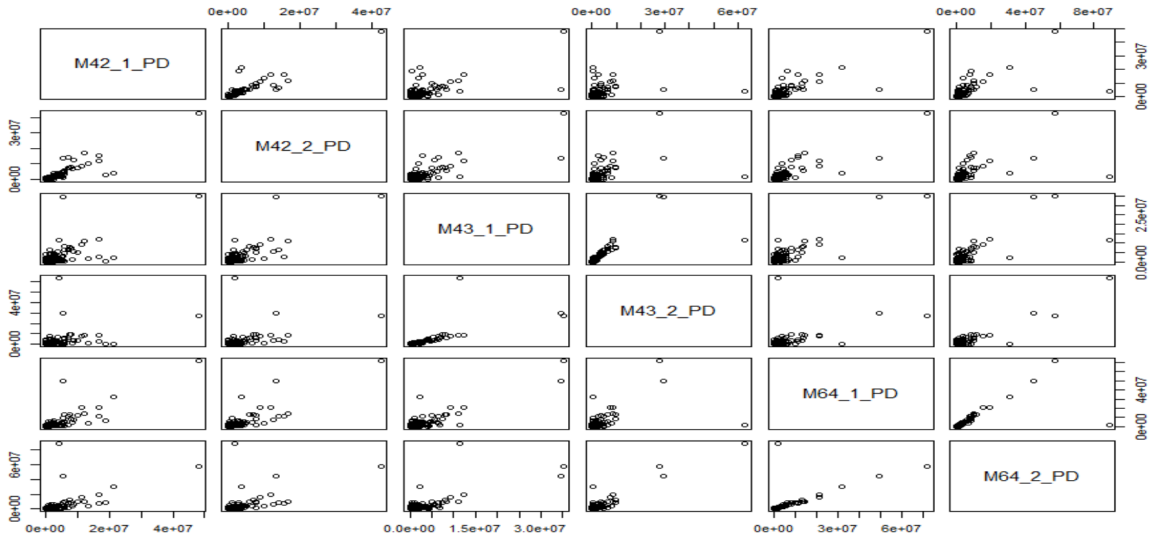
Anexo 1:



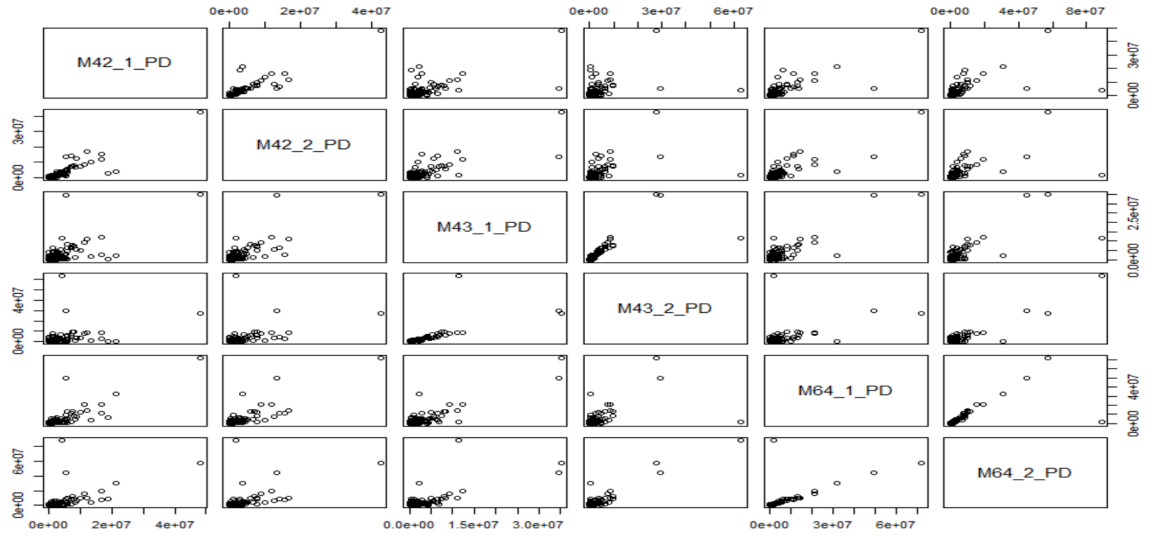

Anexo 2:



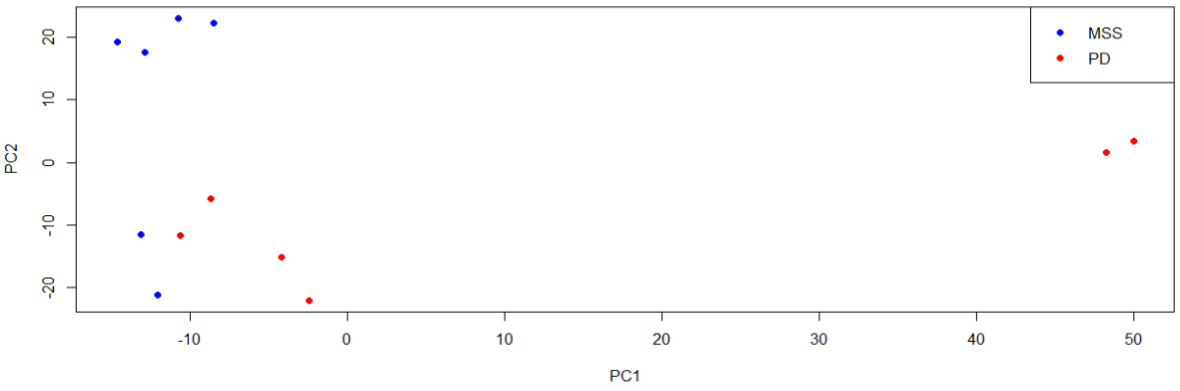
Anexo 3:
MSS:



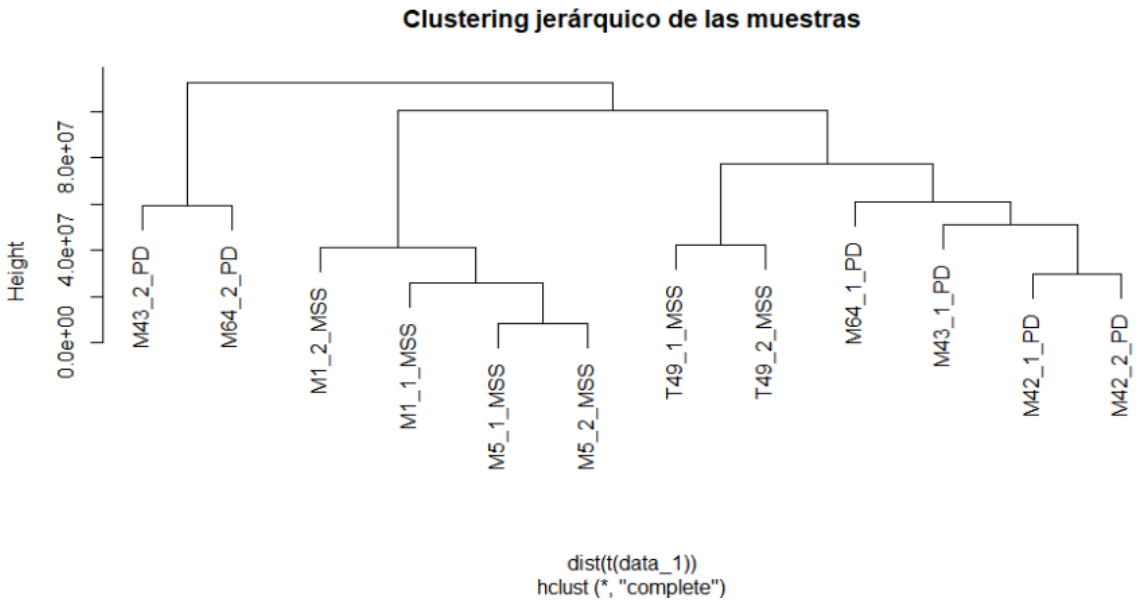

PD:



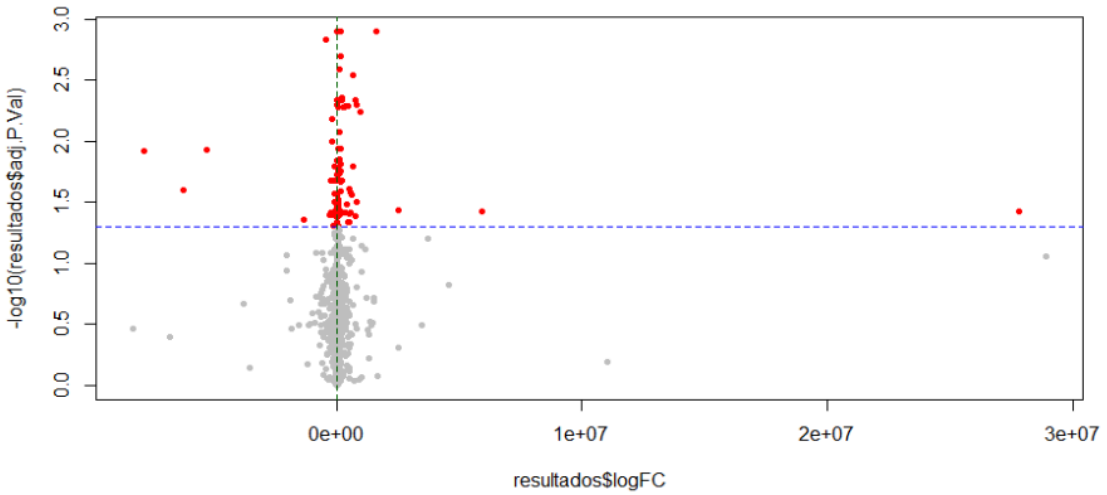

Anexo 4:



Anexo 5:




Anexo 6:




Anexo 7: